Performance des étudiants

Corre, Feteira, Lenoir

Table des matières

1 Introduction et modification des données

À la suite de notre analyse descriptive, nous allons effectuer une analyse factorielle complète sur nos données. Nous avions regroupé nos variables en 4 catégories (renseignements, environnement familial, milieu scolaire et milieu extra-scolaire), nous allons donc réaliser une ACM sur trois des 4 catégories car la catégorie renseignement ne comportait que trop peu de variables et seront mises en variables supplémentaires. Nous allons également, en fonction des résultats obtenus, réaliser une ou plusieurs AFC et classifications. Tout ceci a pour but final de dégager des groupes et sous groupes d'étudiants qui peuvent montrer les facteurs de réussites et d'échecs scolaires au Portugal.

Pour cela, nous avons modifié nos données en créant seulement des variables qualitatives et en adaptant les modalités pour aider à la compréhension.

Deux variables sont le coeur de notre travail, elles seront utilisées dans chacune de nos parties:

— Gradetot : Moyenne de l'étudiant
— Matiere : Matière étudiée

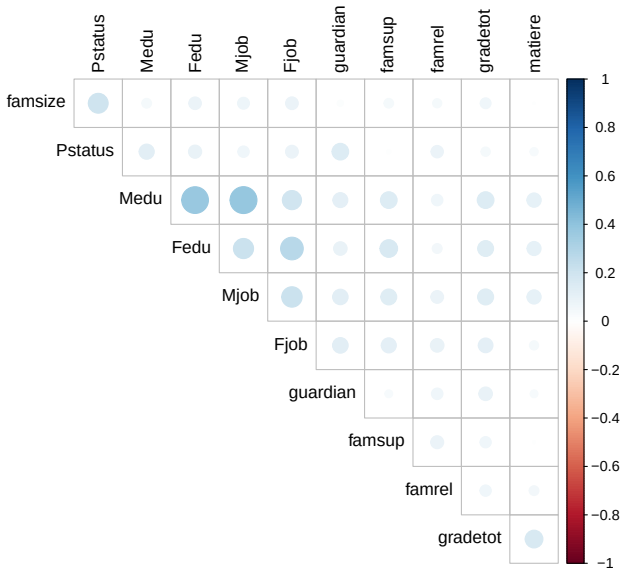
2 L'impact de l'environnement familial

Dans cette partie, les variables liées à l'environnement familial seront étudiées :

— Famsize : Le nombre de personnes dans le foyer
— Pstatut : Le staut conjugal des parents
— Medu et Fedu : Le niveau d'étude des parents
— Mjob et Fjob : La catégorie socioprofessionnelle des parents
— Guardian : Le tuteur de l'étudiant
— Famsup : Soutient à l'éducation
— Famrel : Niveau de qualité des relations familiales

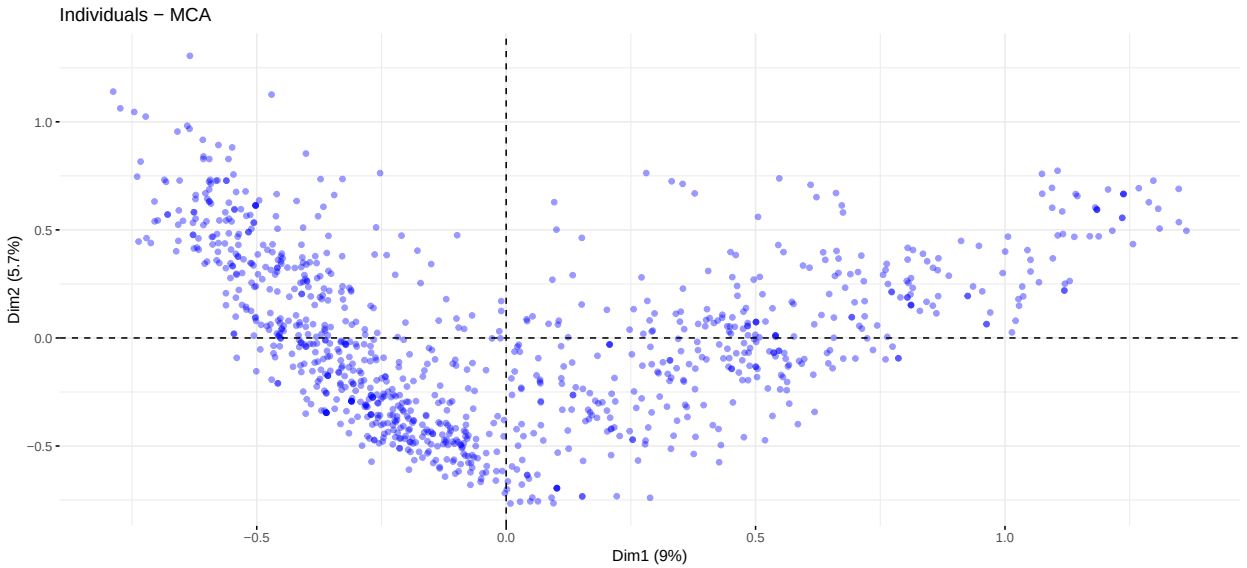
2.1 ACM

2.1.1 Premières modélisations

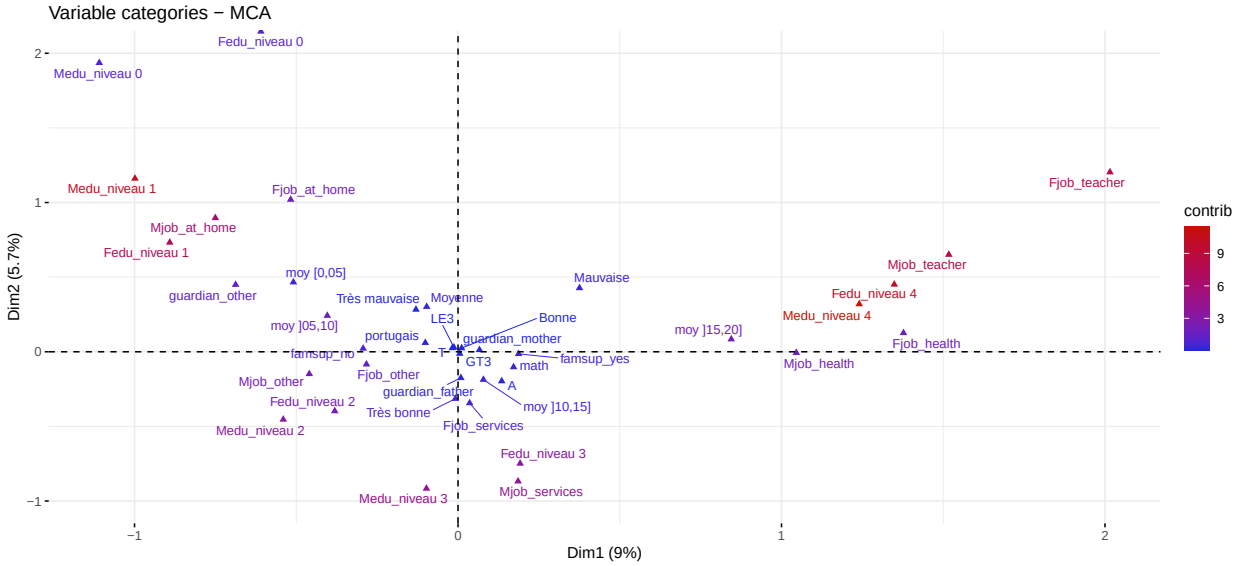


D'après la matrice de corrélation, les variables les plus corrélés entre elles sont Medu avec Fedu et Mjob et Fedu avec Fjob. Ceci semble logique puisque en général, le métier de la mère et du père dépend de son niveau d'étude. De plus, le fait que Medu et Fedu ainsi que Mjob et Fjob soient corrélées peut nous indiquer que les parents se sont probablement rencontrés pendant leurs études ou au travail.

Tout d'abord, jetons un premier coup d'oeil à notre nuage des individus et des variables.

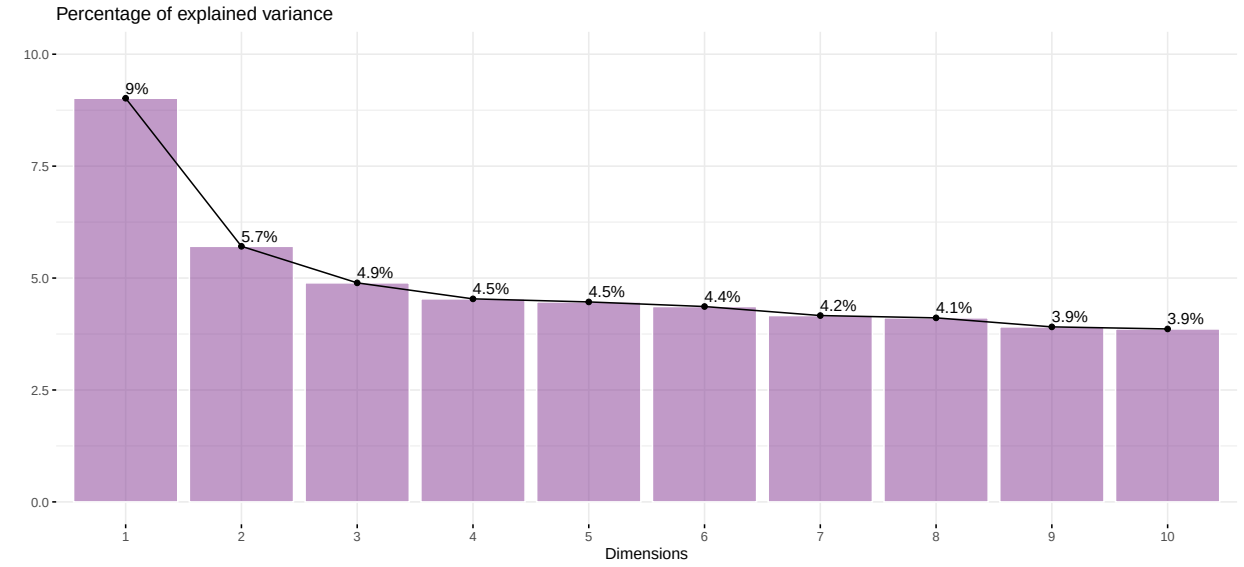


On remarque que le nuage des individus a une forme du type de l'effet Guttman. De plus, on a une très forte concentration des individus du coté gauche de l'axe 2 et une concentration moins importante du côté droit.



Les variables qui semblent contribuer le plus sont Medu, Fedu, Mjob et Fjob. De plus, les modalités concernant le niveau 1 et 4 d'éducation semblent le plus contribuer pour les variables Medu et Fedu ainsi que les modalités "teacher" et "at_home" pour les variables Fjob et Mjob. L'effet Guttman qui apparait dans le nuage des individus est sûrement expliqué par ces variables.

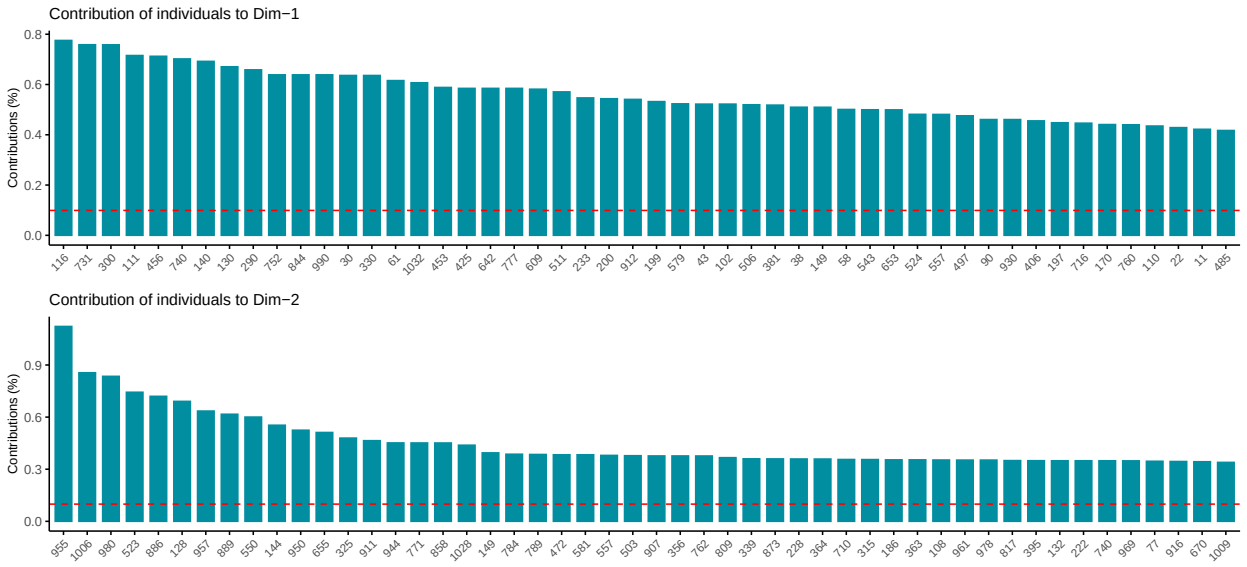
2.1.2 Inerties



On remarque que la dimension 1 à une part de variance expliquée de 9%, pour l'axe 2, nous tombons à 5.7%. Les axes suivants sont très proches. Ici nous choisissons de garder les 2 premiers axes.

2.1.3 Contributions

On regarde quels individus contribuent le plus :



On regarde maintenant les individus qui contribuent le plus à chaque axe :

TABLE 1 – Individus qui contribuent le plus à l axe 1

	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	famsup	famrel	gradetot	matiere
125	GT3	T	niveau 2	niveau 2	other	other	mother	no	Très bonne	moy]05,10]	math
760	GT3	T	niveau 4	niveau 4	health	health	mother	no	Très bonne	moy]15,20]	portugais
321	GT3	A	niveau 4	niveau 3	services	services	mother	yes	Très bonne	moy]10,15]	math
120	GT3	T	niveau 3	niveau 4	other	other	father	no	Moyenne	moy]10,15]	math

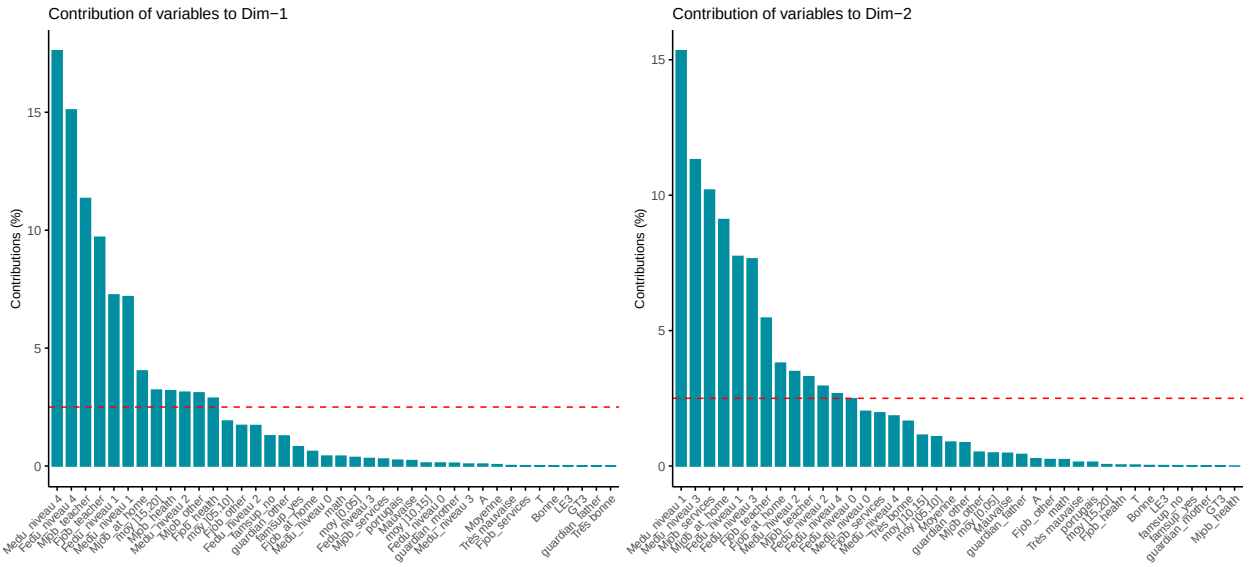
Les étudiants 125, 760 et 321 contribuent le plus à la dimension 1. Une première chose que l'on remarque pour ces étudiants est que les SCP mères et pères sont identiques pour tous les individus. De même, le niveau d'éducation des parents est souvent très proches. Pour ces étudiants nous pouvons également remarquer que les niveaux élevés d'éducation des parents permettent une meilleure moyenne.

TABLE 2 – Individus qui contribuent le plus à l axe 2

	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	famsup	famrel	gradetot	matiere
1000	GT3	T	niveau 1	niveau 1	at_home	services	mother	no	Bonne	moy]05,10]	portugais
1041	LE3	T	niveau 3	niveau 1	teacher	services	mother	yes	Bonne	moy]15,20]	portugais
1015	GT3	T	niveau 3	niveau 3	services	services	mother	yes	Bonne	moy]10,15]	portugais
548	LE3	T	niveau 2	niveau 2	services	services	father	yes	Mauvaise	moy]10,15]	portugais

Pour la dimension 2, l'individu 1000 est celui qui contribue le plus, ensuite nous avons les étudiants 1041 et 1015 qui contribuent quasiment de la même manière. Tout les étudiants ont leurs parents qui vivent ensemble, un père qui travaille dans les services et ont tous passé le portugais. On remarque également que comme pour la dimension 1, lorsque les deux parents ont un niveau d'éducation faible, la moyenne de l'élève baisse, ce qui est le cas pour l'étudiant 1000.

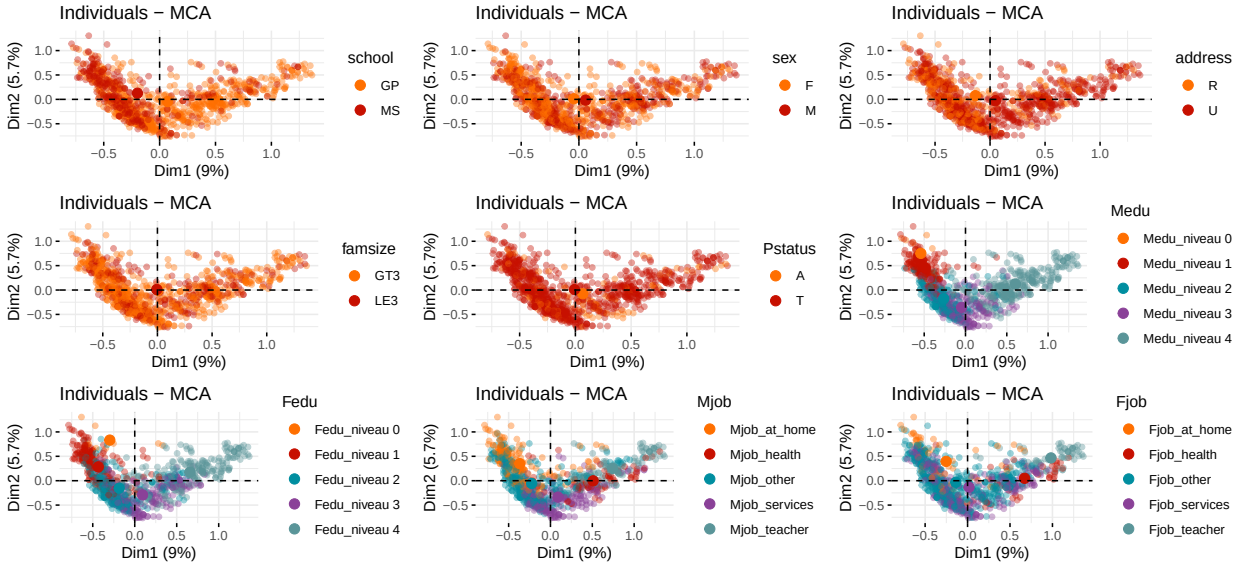
Pour les variables :



Pour l'axe 1, il semble que ce soit le niveau d'éducation qui contribue le plus avec une distinction entre le niveau maximum (4) et le niveau 1. Mais nous pouvons également voir apparaitre Fjob et Mjob teacher ainsi que Mjob et Fjob at home. Donc cela semble séparer les parents qui ont beaucoup d'éducation et qui travaillent dans le secteur de l'éducation et les parents qui ont peu d'éducation et qui travaillent à la maison. L'axe 2 semble également discriminer par le niveau d'éducation, cette fois-ci le niveau 3 apparait ainsi que Fjob et Mjob services. C'est donc la variable education qui donne cette forme au nuage des individus. En effet, l'axe 1 sépare niveau 1 et niveau 4 donc niveau faible et élevé alors que l'axe 2 sépare niveau 1 et niveau 3.

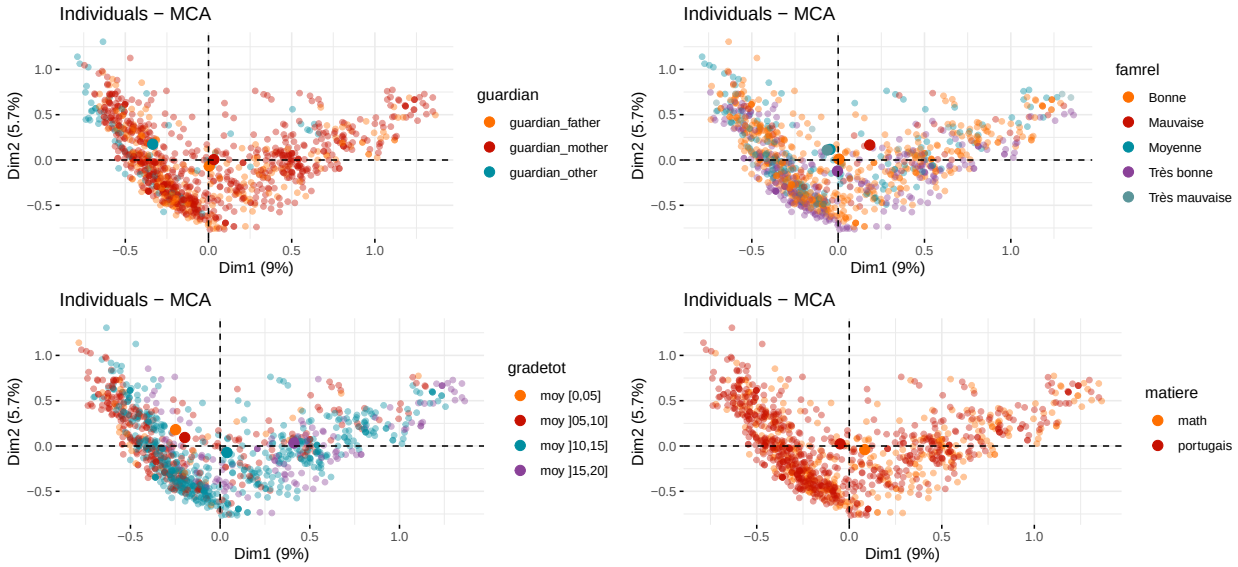
2.1.4 Nuage des individus

Pour les axes 1 et 2:



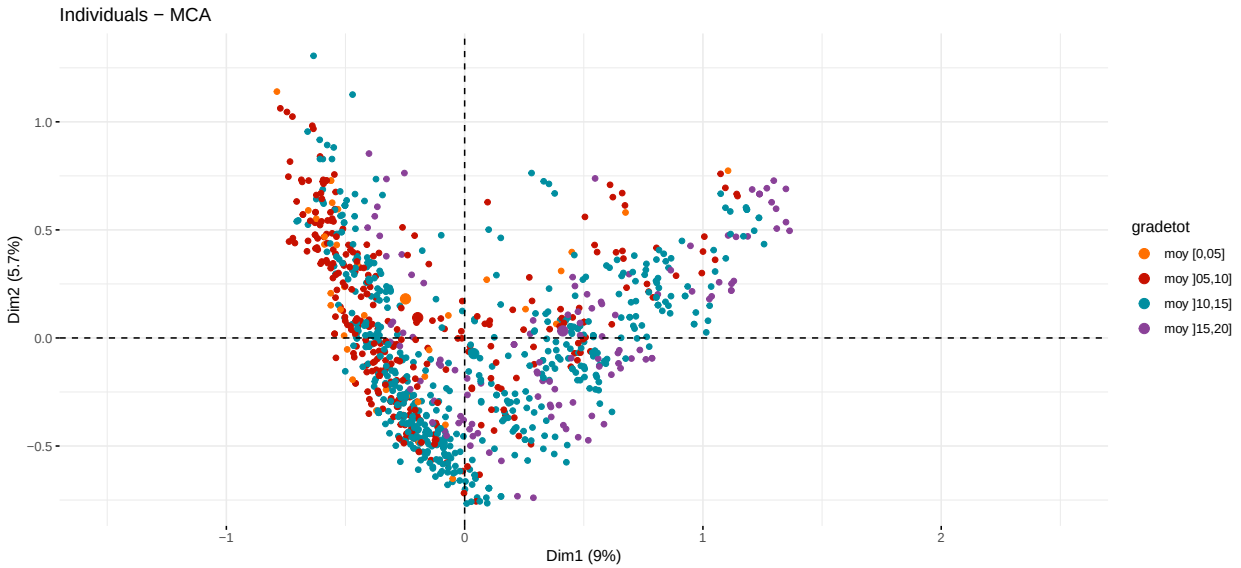
Sur ce graphique, on remarque bien que les variables Medu et Fedu se répartissent du niveau d'éducation le plus faible au niveau le plus élevé sur l'axe 1. Mjob et Fjob semblent se répartir un peu de la même manière, ce qui est logique puisque dans la matrice de corrélation, nous avions une corrélation positive entre ces 4 variables.

Pour la variable school, l'école Gabriel Pereira qui semblait se situer à la ville dans l'analyse descriptive à beaucoup d'individu situés sur la droite, donc des individus dont les parents ont un niveau d'éducation élevé. De plus, pour l'école Mousinho da Silveira, les étudiants sont situés sur la gauche donc ce sont des étudiants avec des parents qui ont un niveau plus faible d'éducation.



Pour toutes les autres variables de notre analyse, nous ne retrouvons pas de résultats très clairs et précis. Les autres variables semblent être expliquées dans d'autres dimensions.

Pour finir, nous allons regarder plus en détail la variable gradetot qui donne la moyenne de l'étudiant en fonction de l'explication des axes déterminés précèdemment.



Sur les axes 1 et 2, il y a une tendance qui se dégage, les très bonnes notes sont plutôt du côté droit, ce qui signifie que les meilleures notes sont plutôt pour les personnes dont le niveau d'éducation des parents est élevé (supérieur au niveau 3). Et au contraire beaucoup de points rouges et oranges sont sur la gauche, ce qui signifie que le niveau d'éducation des parents est faible et les notes sont plutôt mauvaises.

En conclusion de cette ACM, seuls les axes 1 et 2 sont vraiment intéressants à regarder et à analyser. Ces axes montrent que plus le niveau d'éducation de la mère ou du père est élevé, plus la CSP associée semble être dans les secteurs de l'éducation et de la santé. Ceci est logique puisqu'il faut faire des études pour travailler dans ces secteurs. Au contraire ceux avec moins d'éducation travaillent plus à la maison ou dans le secteur administratif et du service.

Pour les notes des étudiants on peut en conclure que plus le niveau d'éducation est élevé plus les notes semblent bonnes. Et donc, par correspondance les enfants de medecins et de professeurs réussissent mieux.

Il serait donc intéressant de faire une AFC sur les variables Medu, Fedu, Mjob et Fjob en lien avec les notes pour confirmer les résultats obtenus.

2.2 AFC CSP mère-notes

Tout d'abord, nous avons réalisé plusieurs AFC sur les variables Medu, Fedu, Mjob et Fjob toutes en relation avec la variable gradetot (Moyenne de l'étudiant). Pour chaque AFC réalisée, les résultats étaient similaires. Nous avons gardé cette AFC car elle semblait être la plus intéressante et présentait les résultats les plus clairs.

2.2.1 Tableaux de données

Voici le tableau de contingence sur lequel nous allons travailler :

Table 3 – Repartition des etudiants selon la moyenne et les CSP de la mere

	at_home	health	other	services	teacher
entre 0 et 4	8	0	7	5	2
entre 4 et 8	21	1	41	21	7
entre 8 et 12	102	24	173	87	53
entre 12 et 14	34	20	103	57	25
entre 14 et 16	15	22	45	31	29
entre 16 et 18	6	8	16	21	8
entre 18 et 20	3	2	2	4	6

On remarque dans ce tableau qu'il y a une surreprésentation pour les notes entre 8 et 16.

2.2.2 Test khi-deux

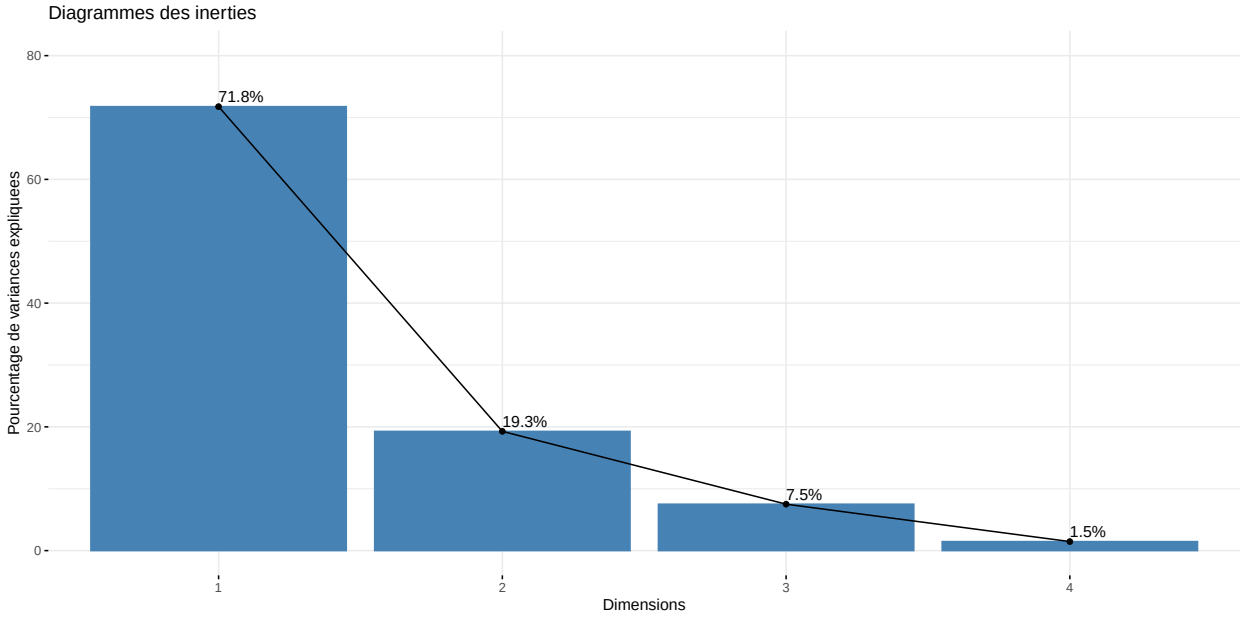
TABLE 4 – Test du Khi-deux

	statistic	parameter	p.value
X-squared	75.61732	24	2.99e-07

La p-value est très faible, on rejette l'hypothèse d'indépendance, les notes et les CSP sont liées, on peut faire une AFC.

2.2.3 Inerties

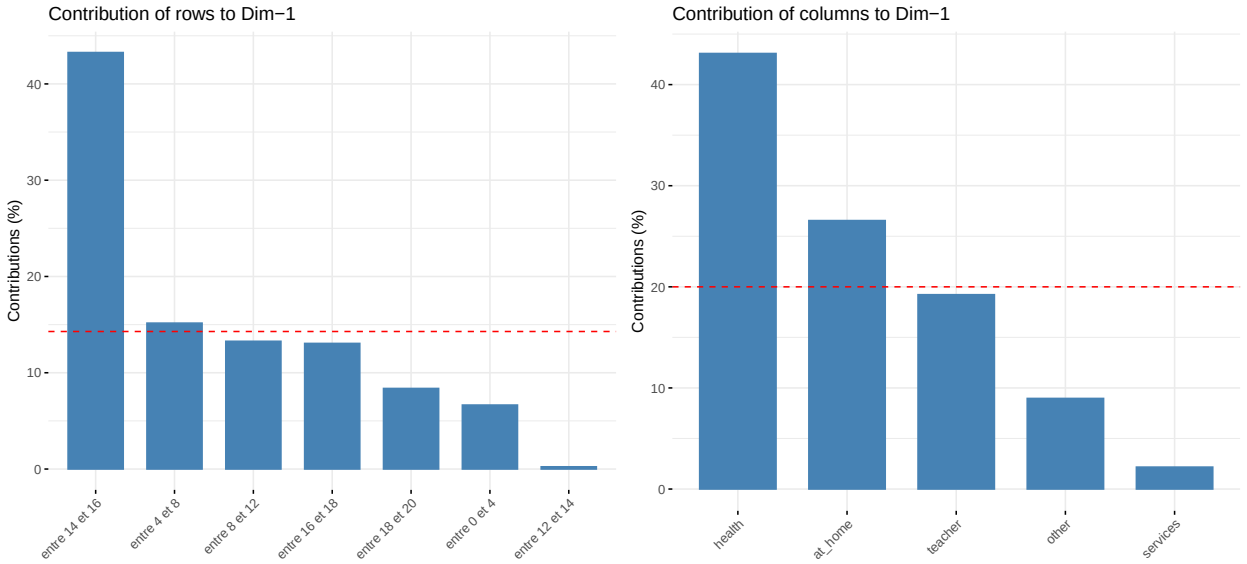
Tout d'abord voici un graphique qui représente les parts d'inertie cumulées :



L'axe 1 explique 72% de la liaison entre les modalités et les deux premiers axes expliquent 91% de la variance. On garde donc deux axes pour effectuer notre AFC.

2.2.4 Contributions

En ce qui concerne les contributions, voici un graphique qui permet d'obtenir les contributions pour la dimension 1 pour les profils lignes et colonnes:



Pour les profils lignes, on remarque que les contributions aux deux premiers axes séparent les bonnes notes des mauvaises notes.

Pour les profils colonnes, on remarque que les personnes travaillant dans les secteurs de la santé et travaillant à la maison sont celles qui contribuent le plus pour les CSP.

En conclusion, on peut dire que les bonnes et mauvaises notes expliquent principalement l'axe 1.

Voici un tableau des contributions qui résument les résultats obtenus graphiquement.

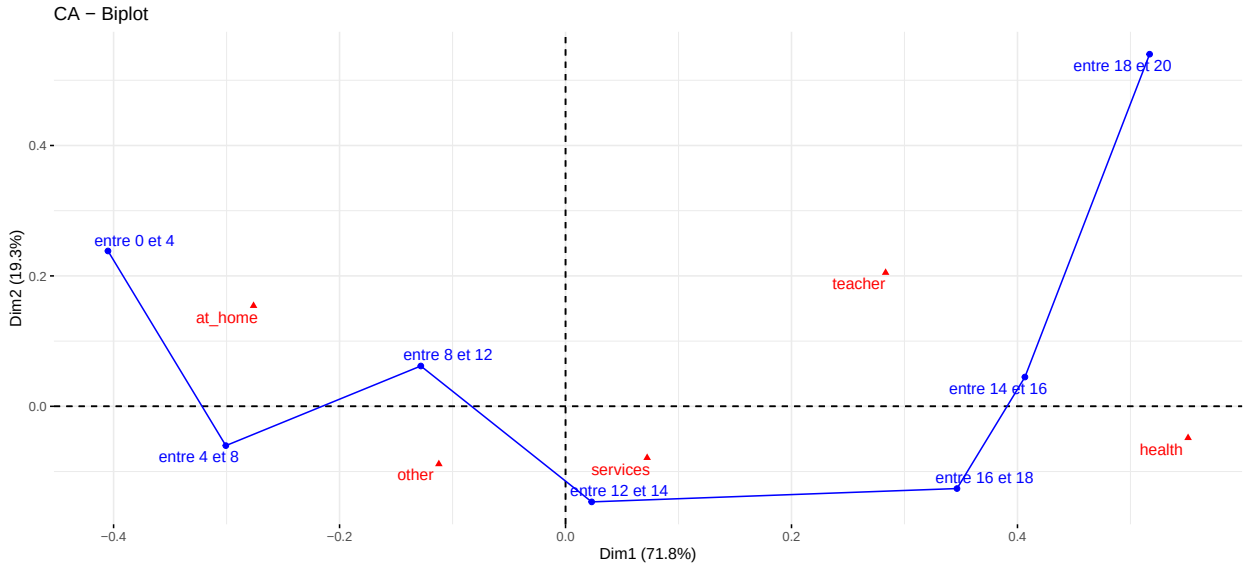
TABLE 5 – Contributions pour chaque axes des profils lignes

	Dim 1	Dim 2	Dim 3	Dim 4
entre 0 et 4	6.65	8.57	14.81	11.30
entre 4 et 8	15.16	2.27	2.19	28.32
entre 8 et 12	13.27	11.52	2.37	7.98
entre 12 et 14	0.24	35.22	2.94	4.31
entre 14 et 16	43.26	1.97	12.18	3.09
entre 16 et 18	13.05	6.44	62.02	3.14
entre 18 et 20	8.37	34.02	3.49	41.86

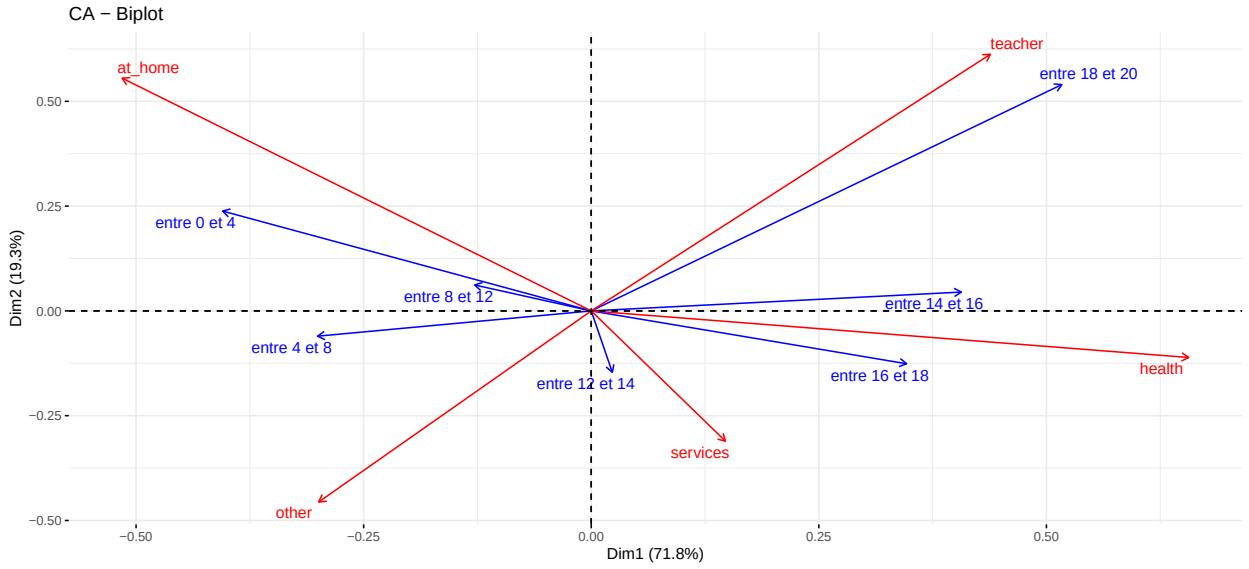
Maintenant que nous avons determiné le nombre d'axe et que nous avons analysé les contributions pour les profils lignes et colonnes, représentons graphiquement les résultats de l'AFC.

2.2.5 Représentations graphiques

Pour finir, représentons les profils lignes et colonnes sur le même graphique :



La représentation de l'axe 1 discrimine les bonnes et mauvaises notes. On voit bien que sur la gauche de l'axe des ordonnées, nous avons les mauvaises notes (entre 0 et 12) et sur la droite, les bonnes notes (entre 12 et 20). On remarque également que les éleves dont la mère travaille dans le secteur de la santé ou de l'éducation ont des meilleures notes que les élèves dont la mère travaille à la maison. Cela renvoie aux observations faites dans l'analyse descriptive.



On ajoute les directions pour confirmer notre analyse sur le graphique précèdent. Sur ce graphique, on arrive à mieux distinguer les CSP, et surtout teacher et health. On peut en conclure que les étudiants qui réussissent le mieux sont les fils dont la mère travaille dans l'éducation et ceux qui réussissent le moins, les enfants dont la mère travaille à la maison.

3 Impact des activités extra-scolaires

Comme vu dans notre analyse descriptive, les variables liées aux activités extra-scolaires sont :

— Romantic : Statut relationnel de l'étudiant
— Internet : Accès à internet à la maison
— Freetime : Niveau de temps libre
— Goout: Niveau de sortie avec des amis
— Alcool : Niveau de consommation d'alcool

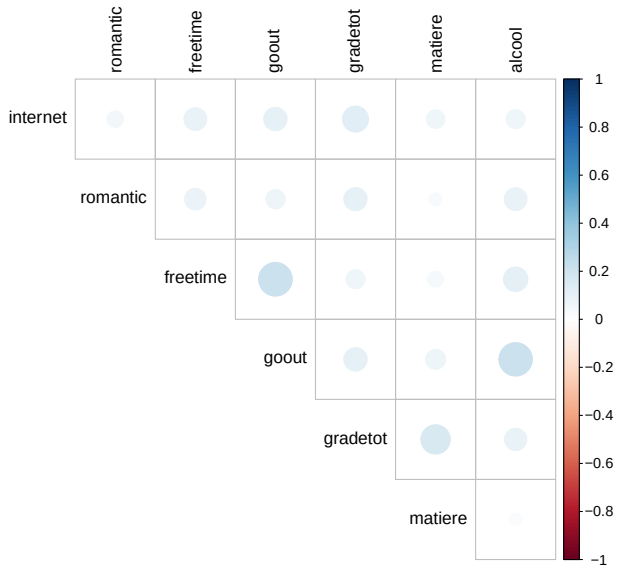
3.1 ACM

Dans cette partie, nous avons choisi d'interpréter les variables liées à l'extra scolaire : goout, freetime, romantic, internet, Dalc et Walc.

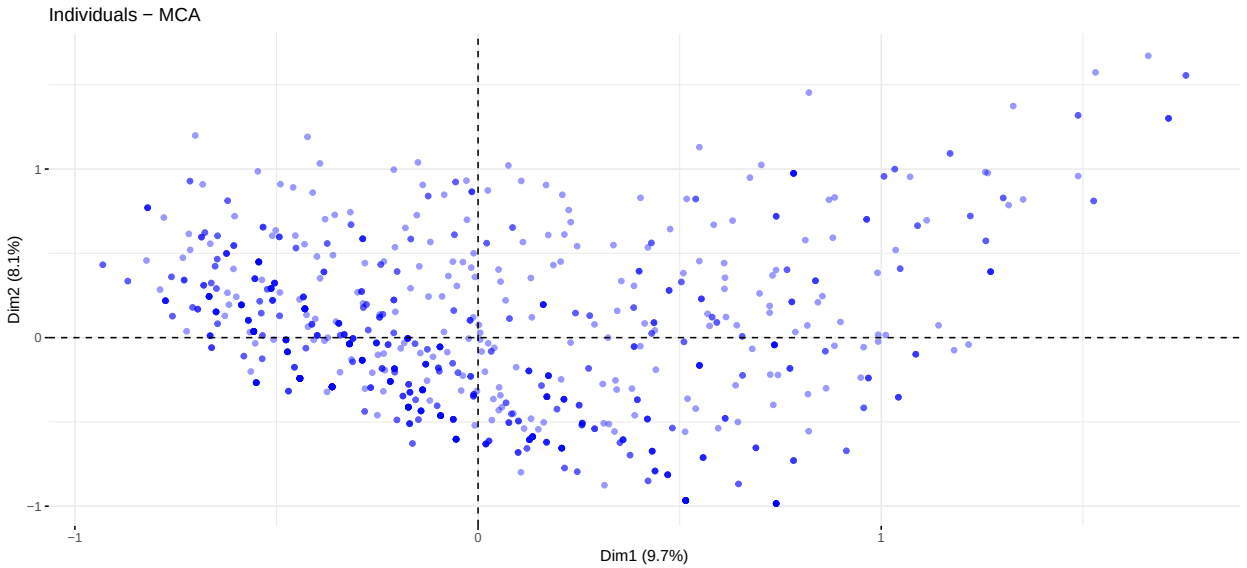
On a pu constater que les deux variables liées à l'alcool étaient corrélées. Le coefficient de corrélation entre ces deux variables est le suivant :

— 0.61 pour les élèves ayant étudié le portugais.
— 0.66 pour ceux ayant étudié les maths.

Nous avons donc créé une nouvelle variable : "alcool" qui représente la consommation d'alcool la semaine et le week-end.

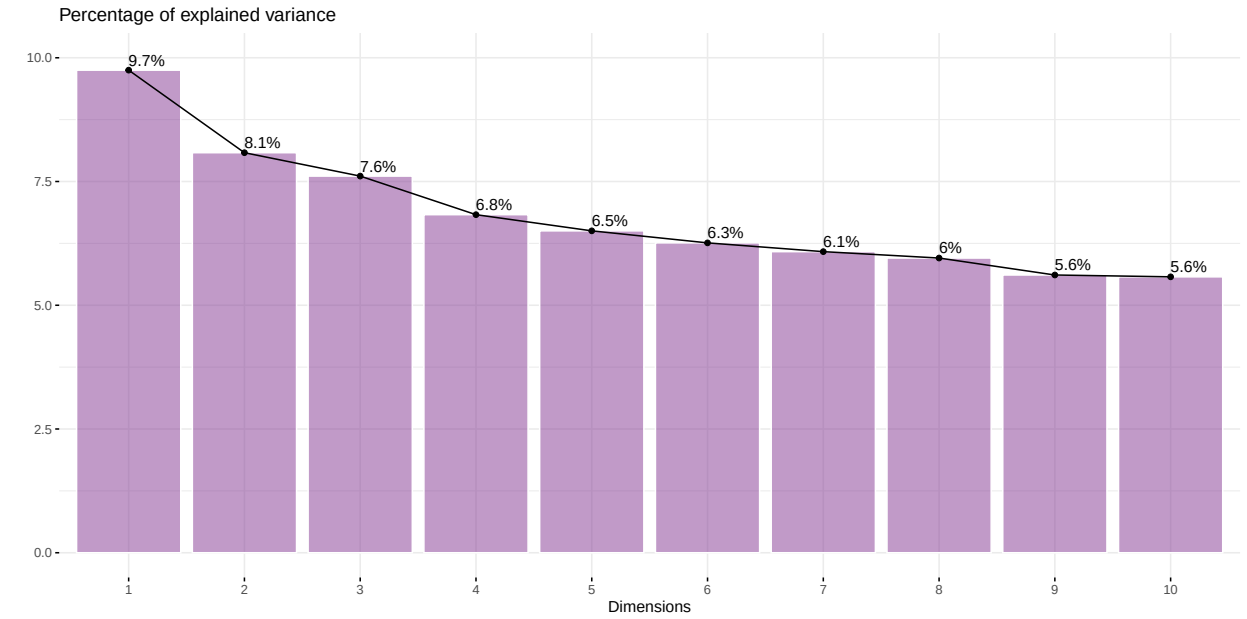


La variable qui représente les sorties entre amis est corrélée avec deux autres variables : le niveau de temps libre et la consommation d'alcool.



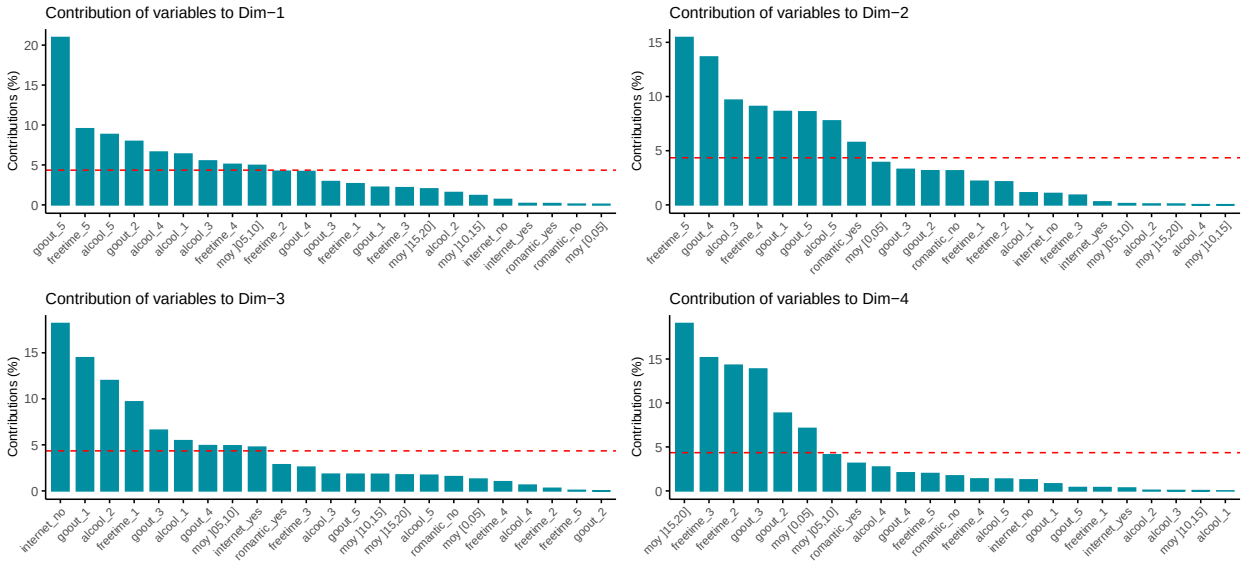
Pour les deux premiers axes, on voit que la représentation des individus à une forme en v. La plupart des individus sont situés à gauche et quelques individus sont situés en haut à droite.

3.1.1 Inertie



Nous décidons de garder nos 4 premiers axes, ce qui correspond à une inertie de 32.3%

3.1.2 Variables



Sur les deux premiers axes, les nombreuses sorties entre amis et le temps libre important contribuent beaucoup à ces deux axes :

— Sur l'axe 1, goout_5 contribue à plus de 20%.
— Freetime_5 et goout_4 contribuent à eux deux presque 30% de l'axe 2.

Sur le troisième axe, la modalité "internet_no" contribue le plus avec également les modalités les plus basses des sorties entre amis, de la consommation d'alcool et du temps libre.

Sur notre quatrième axe, nous voyons que les moyennes élevées contribuent à presque 20%. Mais aussi que les modalités moyennes des variables du temps libre et des sorties entre amis y contribuent pour plus de 10% chacune.

3.1.3 Individus

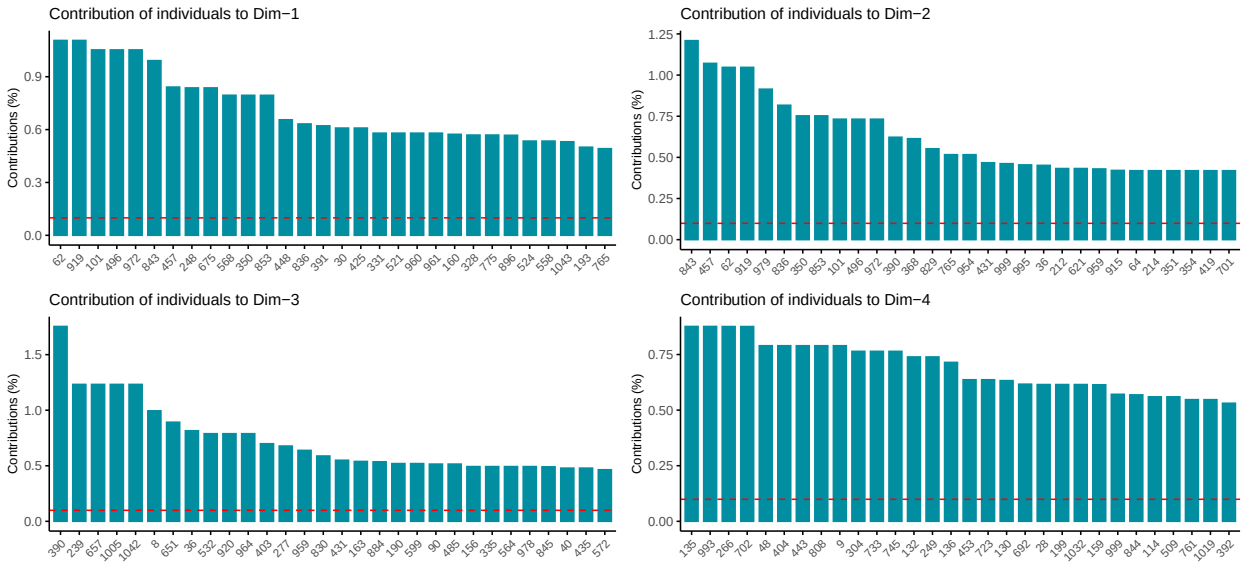


Table 6 – Individus à forte contribution sur l'axe 1

	internet	romantic	freetime	goout	gradetot	matiere	alcool
62	yes	yes	5	5	moy]05,10]	math	5
919	yes	yes	5	5	moy]05,10]	portugais	5
101	yes	no	5	5	moy]05,10]	math	5
496	yes	no	5	5	moy]05,10]	portugais	5
972	yes	no	5	5	moy]05,10]	portugais	5

Les individus qui contribuent le plus à l'axe 1 sont presques identiques sur de nombreux points : individus qui ont beaucoup de temps libre, qui sortent beaucoup, qui ont une consommation d'alcool dans la semaine comme le week-end et donc à cause de cela, des moyennes inférieures à 10.

Table 7 – Individus à forte contribution sur l'axe 2

	internet	romantic	freetime	goout	gradetot	matiere	alcool
843	no	yes	5	5	moy]05,10]	portugais	5
457	yes	yes	5	5	moy]10,15]	portugais	5
62	yes	yes	5	5	moy]05,10]	math	5
919	yes	yes	5	5	moy]05,10]	portugais	5

Les individus qui contribuent à l'axe 2 sont pour la plupart identiques à ceux qui contribuent à l'axe 1 : beaucoup de temps libre, beaucoup de sortie et beaucoup d'alcool.

Table 8 – Individus à forte contribution sur l'axe 3

	internet	romantic	freetime	goout	gradetot	matiere	alcool
390	no	no	1	1	moy [0,05]	math	1
239	no	no	1	1	moy]10,15]	math	1
657	no	no	1	1	moy]10,15]	portugais	1
1005	no	no	1	1	moy]10,15]	portugais	1
1042	no	no	1	1	moy]10,15]	portugais	1

Sur notre axe 3, les individus sont l'opposé de ceux de l'axe 1 et 2 : ils n'ont pas internet, ne sont pas en couple, n'ont pas beaucoup de temps libre, ne sortent pas beaucoup et ne boivent pas d'alcool. À part l'étudiant 390 qui a une moyenne très faible, les autres ont des moyennes plus élevées que les individus des deux premiers axes.

Table 9 – Individus à forte contribution sur l'axe 4

	internet	romantic	freetime	goout	gradetot	matiere	alcool
135	no	yes	3	3	moy [0,05]	math	1
993	no	yes	3	3	moy [0,05]	portugais	1
266	yes	no	2	5	moy]15,20]	math	4
702	yes	no	2	5	moy]15,20]	portugais	4

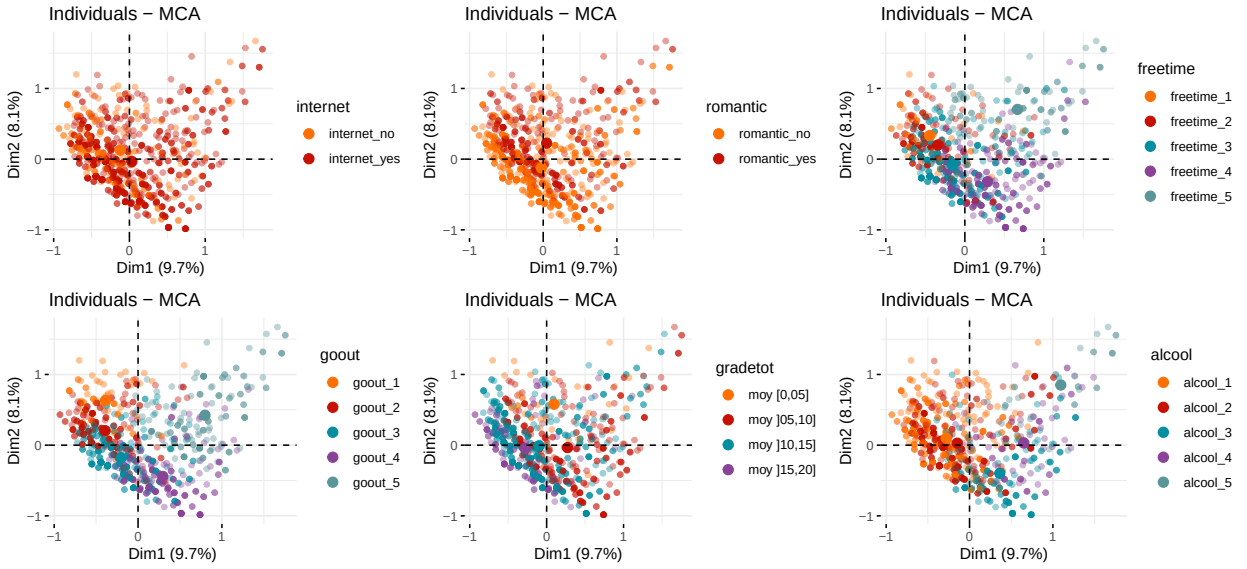
Ici, on distingue deux groupes d'individus qui sont très différents et qui contribuent à l'axe 4 :

— Ceux qui ont une moyenne très basse (en dessous de 5), qui ne boivent pas d'alcool, qui sortent à une fréquence moyenne, qui n'ont pas internet et qui sont en couple.

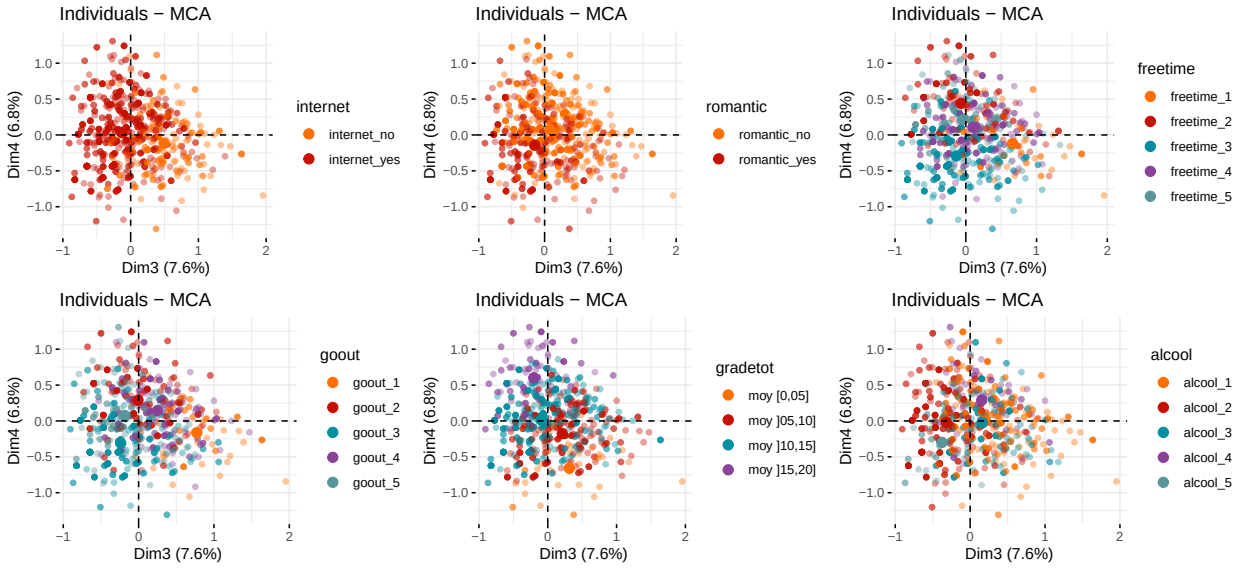
— Ceux qui ont une moyenne très élevée (au dessus de 15), qui boivent plus que la moyenne, sortent beaucoup, qui ont internet et sont en couple.

Ces deux groupes sont biens distincts mais les individus qui font partie du même groupe sont exactement pareils.

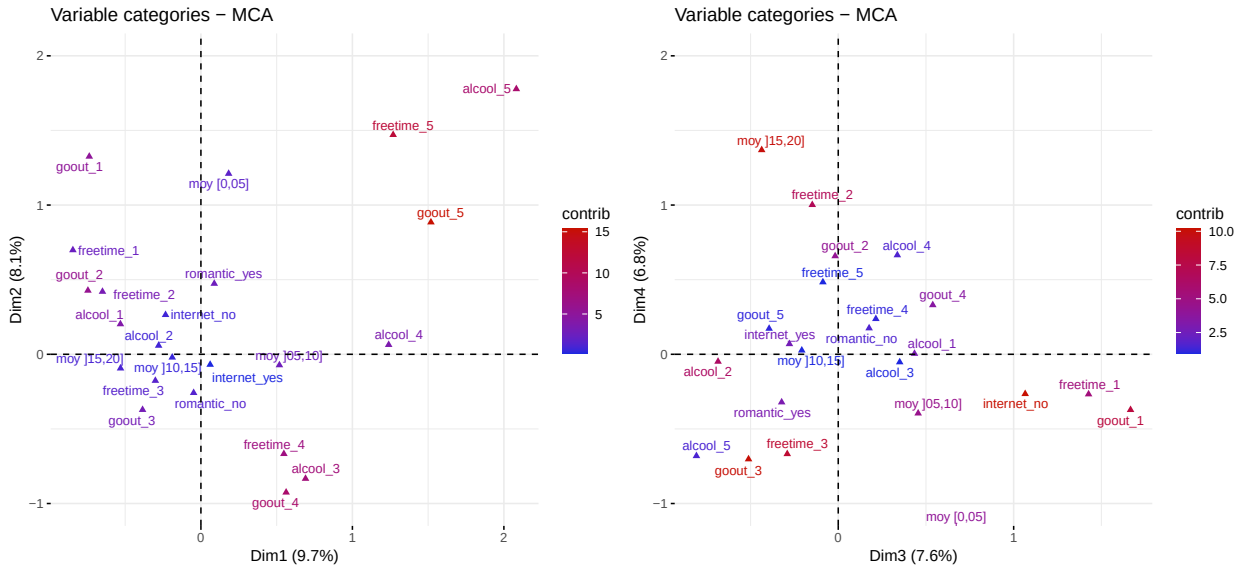
3.1.4 Analyse des individus et des variables.



Sur la représentation des axes 1 et 2, on constate que chaque modalité des variables goout, freetime et alcool est bien représentée distinctement des autres. La modalité supérieure de ces 3 variables est située en haut à droite de leur graphique. Sur la gauche, les modalités inférieures sont représentées. L'axe 1 représente donc la tendance de ces trois variables. Pour les variables internet, romantic et gradetot, il est difficile de trouver une tendance générale.



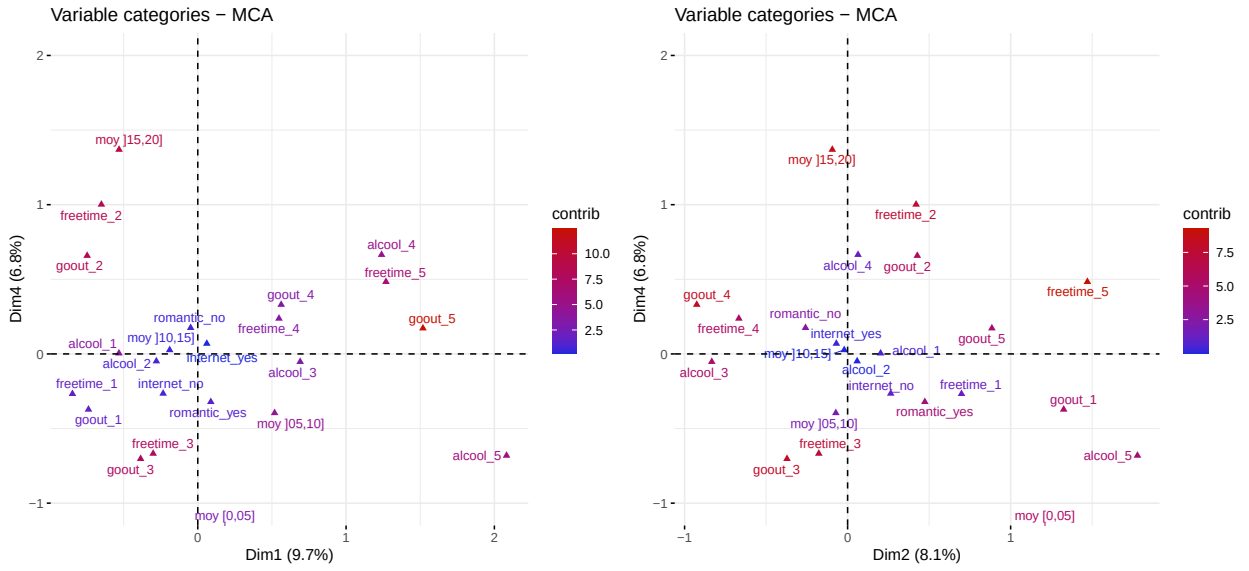
Sur ces deux axes, deux variables sont mieux représentées que les autres. Les individus qui ont internet sont situés sur la gauche et ceux qui ne l'ont pas, sur la droite. Les individus avec une moyenne élévée sont sur la partie haute du graphique et plus la moyenne descend, plus ils se placent vers le bas. Toutes les autres modalités des variables sont assez dispersées sur ces deux axes.



Comme vu dans la matrice de corrélation, on voit que alcool, freetime et goout sont liées. Sur l'axe 1 et 2, la modalité supérieure de ces 3 variables contribue beaucoup à ces deux axes. Ce qui signifie que plus un étudiant a du temps libre, plus il va sortir avec ses amis et plus il va consommer d'alcool.

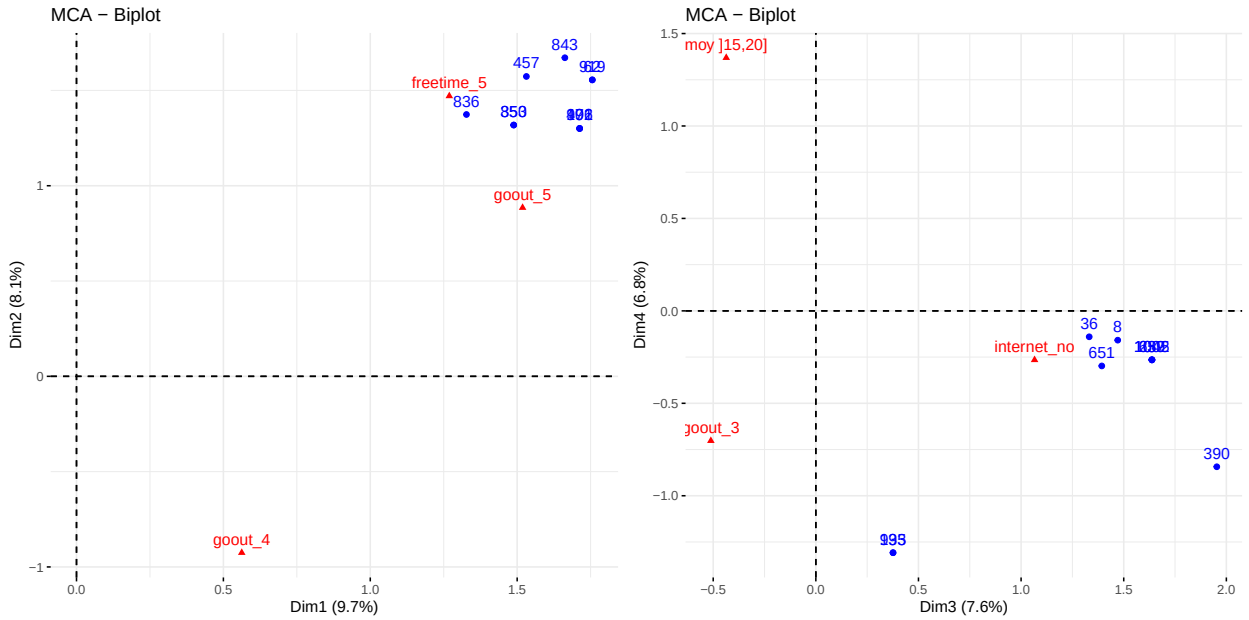
Le deuxième graphique signifie que si un étudiant ne possède pas internet, il aura en moyenne des moins bonnes notes.

Dans les prochains graphiques, nous allons représenter les deux premiers axes avec l'axe 4 qui représente les bonnes moyennes. À côté de la contribution pour ces axes, on décide de faire une représentation des individus pour la moyenne avec les mêmes axes choisis.



Nous constatons que plus un individu va avoir de temps libre ou sortir avec ses amis, plus sa moyenne sera basse. Comme vu précédemment, si un individu a beaucoup de temps libre, il va donc beaucoup sortir avec ses amis et donc sa moyenne sera encore plus basse.

3.1.5 Recapitulatif



Nous avons décidé de représenter les 3 variables et les 10 individus qui contribuent le plus à leurs axes. Sur les deux premiers axes, on constate bien que se sont les individus en haut à droite qui contribuent le plus où ils ont comme spécificité d'avoir beaucoup de temps libre et de beaucoup sortir avec leur amis. Sur les deux axes suivants, les individus qui contribuent le plus sont situés dans le quart en bas à droite. Ces individus font contribuer les modalités "internet_no" et sont à l'opposé des moyennes élevées, ce qui renforce cette modalité.

3.2 AFC alcool-notes

TABLE 10 – p-value des tests du khi-deux pour les variables extra-scolaire

	internet	romantic	freetime	goout	alcool
p-value	0.0006902	0.0128716	0.2599167	0.0008764	0.0065413

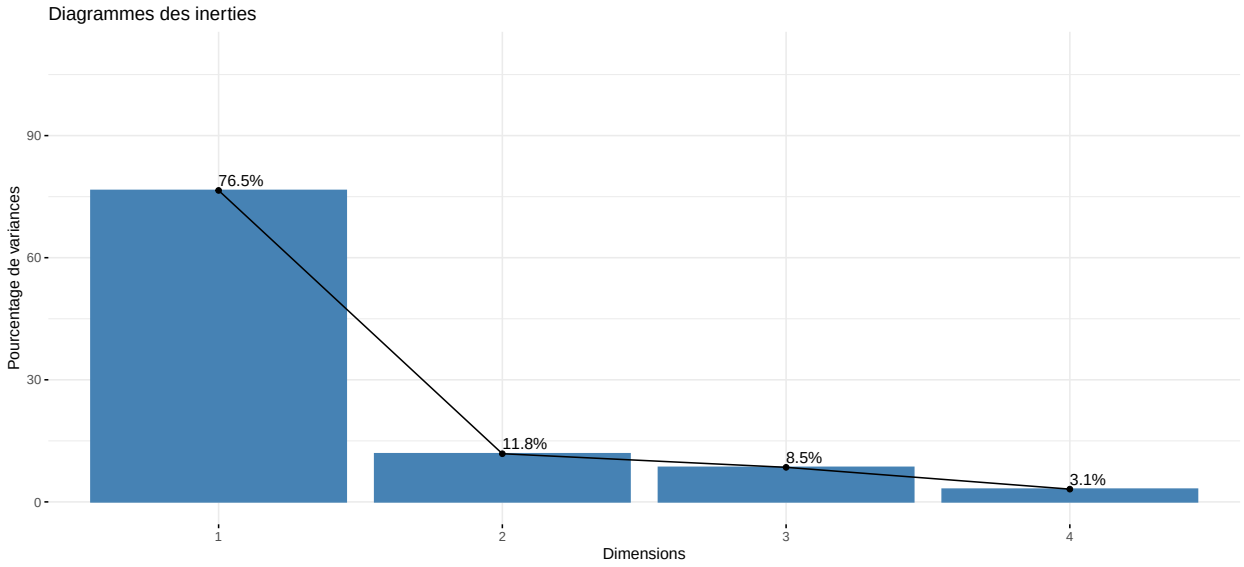
On constate que toutes les p-values sont inférieures à 0.05 sauf la variable freetime donc on peut rejeter l'hypothèse d'indépendance. Nous ne traitons pas la variable romantic et internet car le nombre de modalité est trop faible. Nous décidons de travailler sur la variable "alcool" car dans l'ACM qui précède, on a pu constater que cette variable était importante sur tous les axes mais ne contribuait pas de manière importante.

TABLE 11 – Tableau de contingence

	TP	P	M	E	TE
Inf 8	46	43	27	10	7
Entre 8-10	65	57	58	23	9
Entre 10-12	84	85	42	24	10
Entre 12-14	92	76	41	7	7
Entre 14-16	58	54	18	5	1
Sup 16	32	18	7	3	0

On voit que la consommation très importante d'alcool est une modalité rare.

3.2.1 Inerties



En conservant les deux premiers axes, on obtient 88.37% d'inertie. Le second axe va juste nous permettre une bonne représentation de nos modalités.

3.2.2 Profils lignes-colonnes

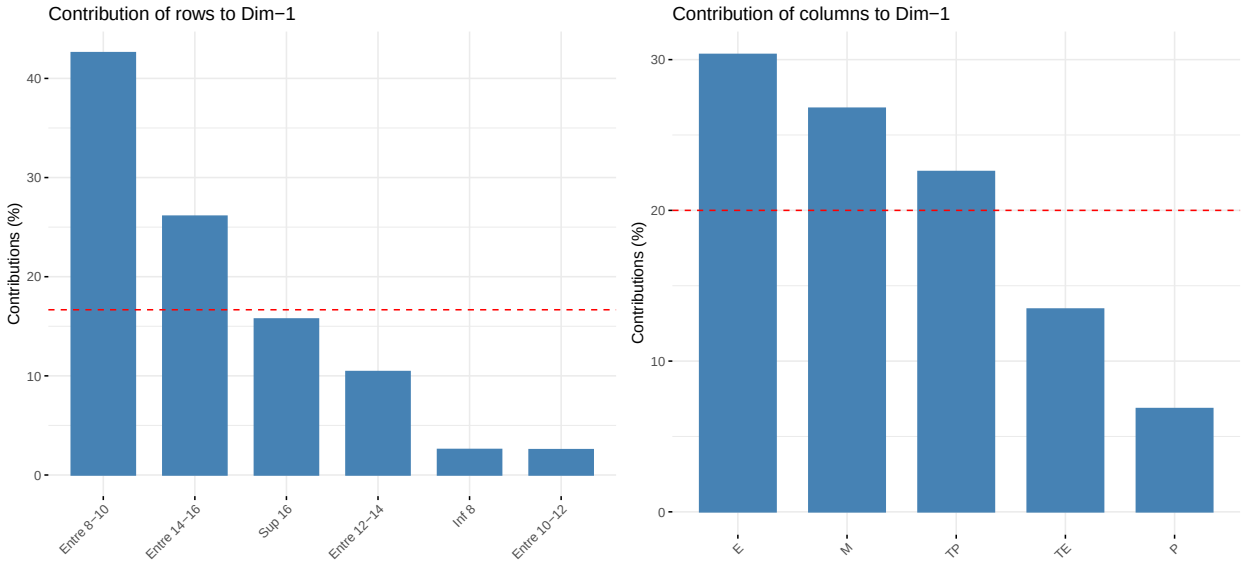


TABLE 12 – Contributions pour chaque axes des profils lignes

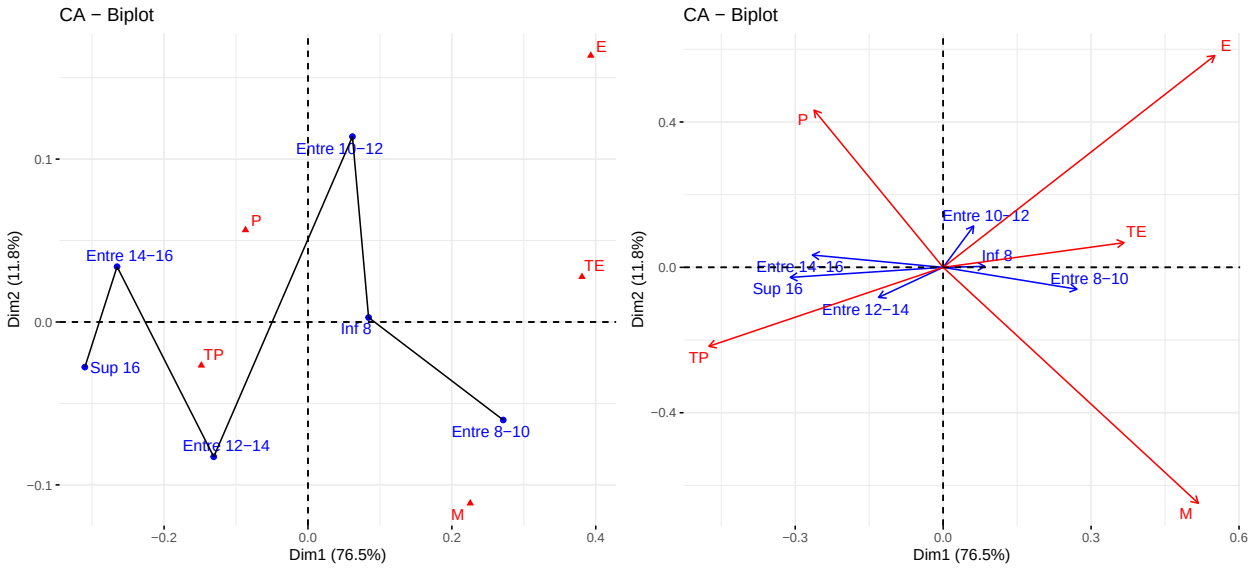
	Dim 1	Dim 2	Dim 3	Dim 4
Inf 8	2.58	0.02	14.62	19.33
Entre 8-10	42.60	13.53	12.12	10.72
Entre 10-12	2.55	55.95	0.03	1.71
Entre 12-14	10.43	26.92	17.05	1.86
Entre 14-16	26.11	2.76	0.00	45.86
Sup 16	15.74	0.81	56.18	20.52

TABLE 13 – Contributions pour chaque axes des profils colonnes

	Dim 1	Dim 2	Dim 3	Dim 4
TP	22.58	4.72	14.98	20.36
P	6.85	18.72	20.70	20.73
M	26.78	42.14	0.03	11.93
E	30.34	33.96	28.38	0.17
TE	13.45	0.46	35.91	46.80

On constate que pour les profils lignes, les individus ayant eu entre 8 et 10 et entre 12 et 16 contribuent à l'axe 1 pour presque 70%. Pour les profils colonnes, la contribution à l'axe 1 est due à trois modalités: très peu, moyen et élevé. Ces 3 modalités contribuent entre 20 et 30% chacune.

3.2.3 Illustrations



Ici, nous pouvons constater que nos analyses passées étaient correctes :

- À gauche, nous avons les moyennes élevées et la consommation faible d'alcool.
- À droite, les moyennes plus faibles et une consommation d'alcool plus élevée.
- Quand la consommation d'alcool est très peu importante, les notes sont meilleures.
- Plus la consommation est élevée, plus les notes sont basses. On peut observer les individus qui ont des moyennes inférieures à 8 ont un vecteur qui se place sur celui des individus qui ont une consommation d'alcool très élevée.

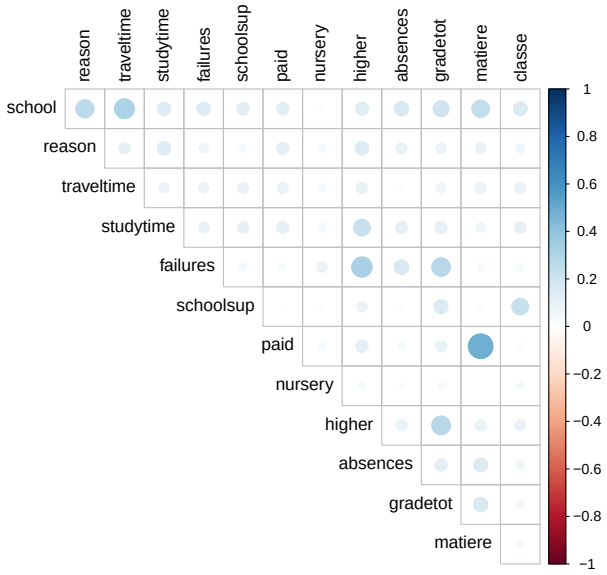
4 Impact du milieu scolaire.

Ici, l'impact du milieu scolaire sera étudié. Les variables liées à ce milieu sont les suivantes :

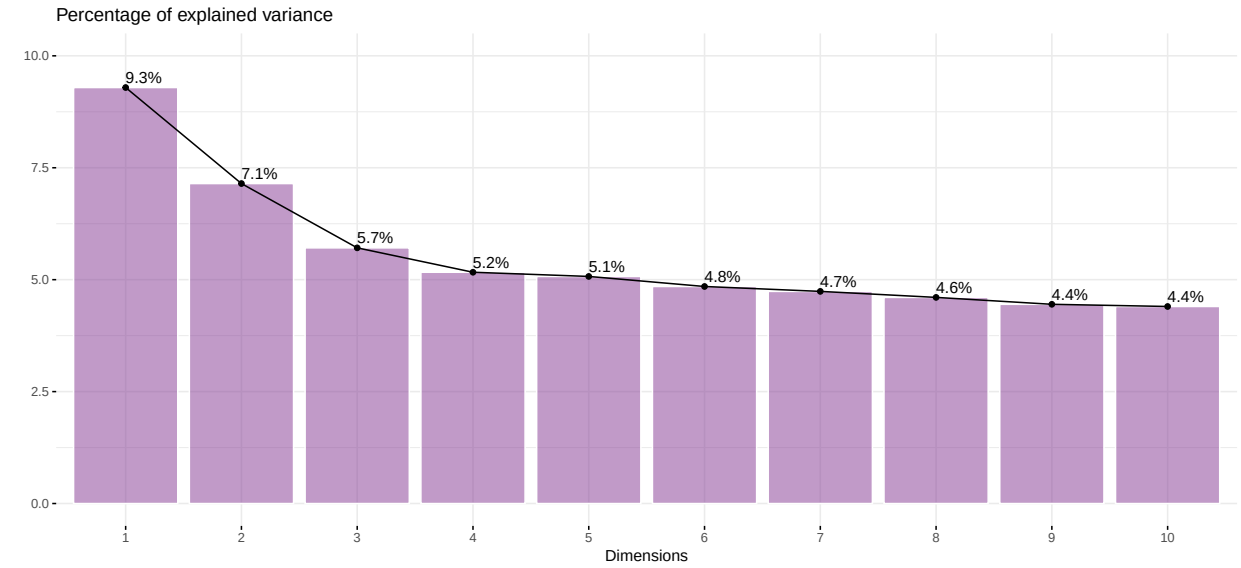
— School : L'école de l'étudiant
— Reason : La raison du choix de l'école
— Traveltime : Temps de trajet pour aller à l'école
— Studytime : Temps de travail par semaine
— Failures : Nombre de redoublement
— Schoolsup : Soutient extra-scolaire
— Paid : Cours supplémentaire payé
— Nursery : L'étudiant a été en maternelle
— Absences : Nombre d'absences
— Classe : Classe de l'étudiant

4.1 ACM

On décide maintenant de faire une ACM sur les variables du milieu scolaire. Pour cela on regarde la matrice de corrélation de ces variables :

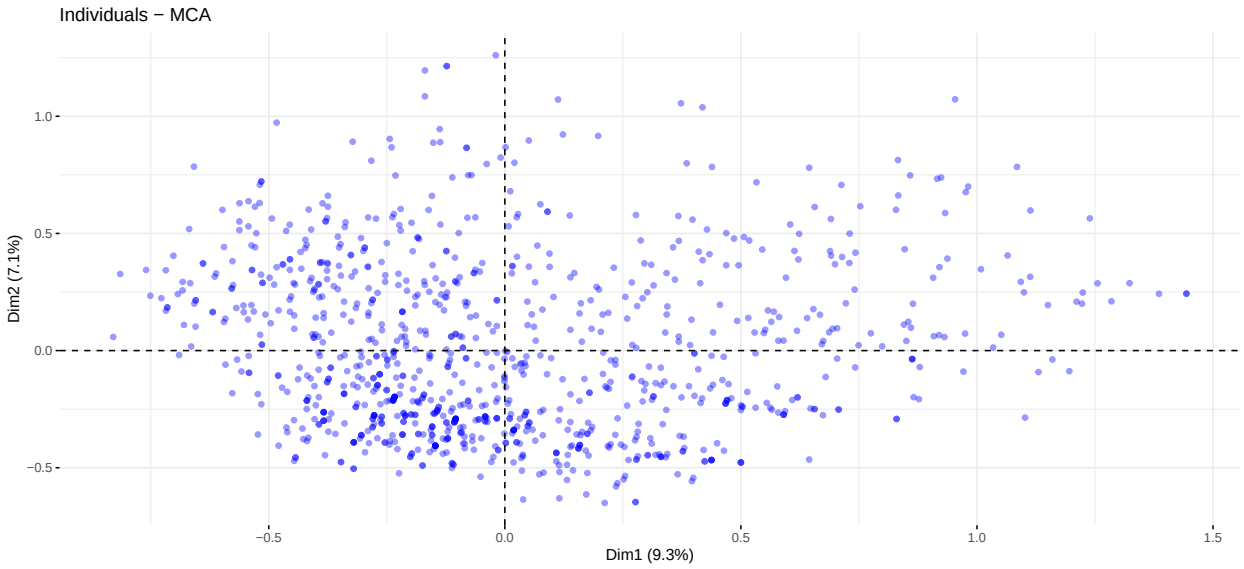


On remarque que paid est très corrélée avec matière. En effet, on sait de notre première analyse que les étudiants prenaient souvent un professeur particulier en mathématiques. De plus, la variable school est corrélé avec traveltime et avec quasiment toutes les autres variables (sauf nursery). Comme nous l'avions vu dans le premier rapport, les étudiants de l'école Gabriel Pereira habitent plutôt relativement proche de l'école (moins de 15 minutes), tandis que les étudiants de l'école Mousinho da Silveira habitent un peu plus loin (entre 15 et 30 minutes) en moyenne. Enfin, la variable failures est corrélée avec higher et gradetot, ce qui nous semble assez logique car moins nos notes sont bonnes, plus l'on redouble et moins l'on a envie de continuer à faire des études (les mauvaises notes sont souvent liées au fait que l'on aime pas l'école).
Nous faisons l'ACM, ce qui nous donne les parts de variance expliquée suivantes :

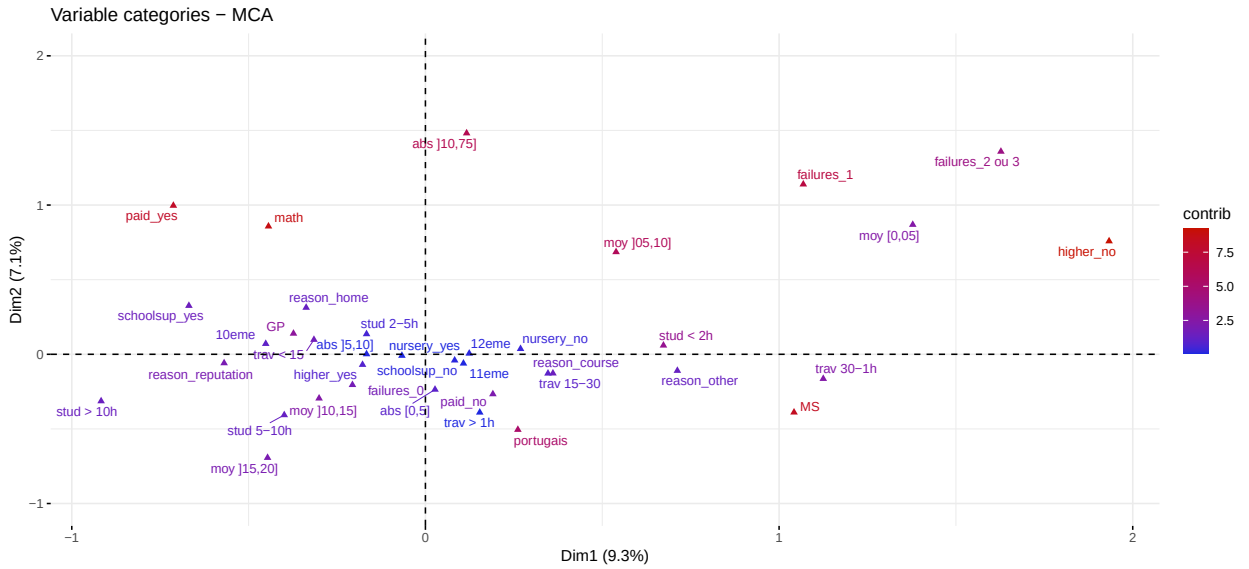


On a 16.4% d'inertie sur les 2 premiers axes.

Voici le graphique des individus sur l'axe 1 et 2 :



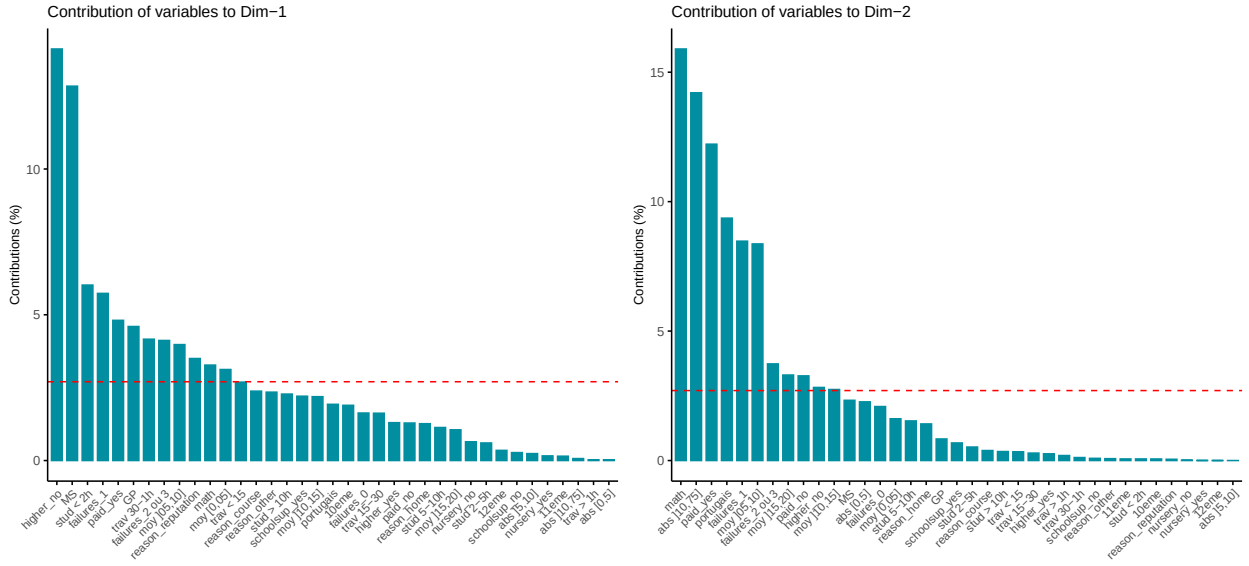
Quelques individus expliquant un peu plus l'axe 1 et 2 mais globalement gros nuage de points. Néanmoins, on observe une forte concentration des points en bas et à gauche.



Les modalités expliquant l'axe 1 et l'axe 2 sont celles qui représentent le fait d'avoir de mauvaises notes tout en ayant une taille d'échantillon plus faible que les autres (c'est pour cela que les points se détachent du centre). On y retrouve donc le fait de ne pas vouloir faire d'études supérieures (higher_no), le fait d'avoir redoublé au moins 1 fois (failures_1 et failures_2 ou 3), le fait d'être absent plus de 10 fois dans l'année (abs]10,75]), le fait de payer un professeur particulier (paid_yes) et la matière mathématiques.
L'axe 1 est également expliqué par les étudiants qui ont un trajet entre 30 minutes et 1 heure (trav 30-1h) et l'école Mousinho da Silveira (MS).

4.1.1 Variables

On étudie les contributions des variables sur les 2 premiers axes car les axes 3 et 4 ne nous apportent aucune information :

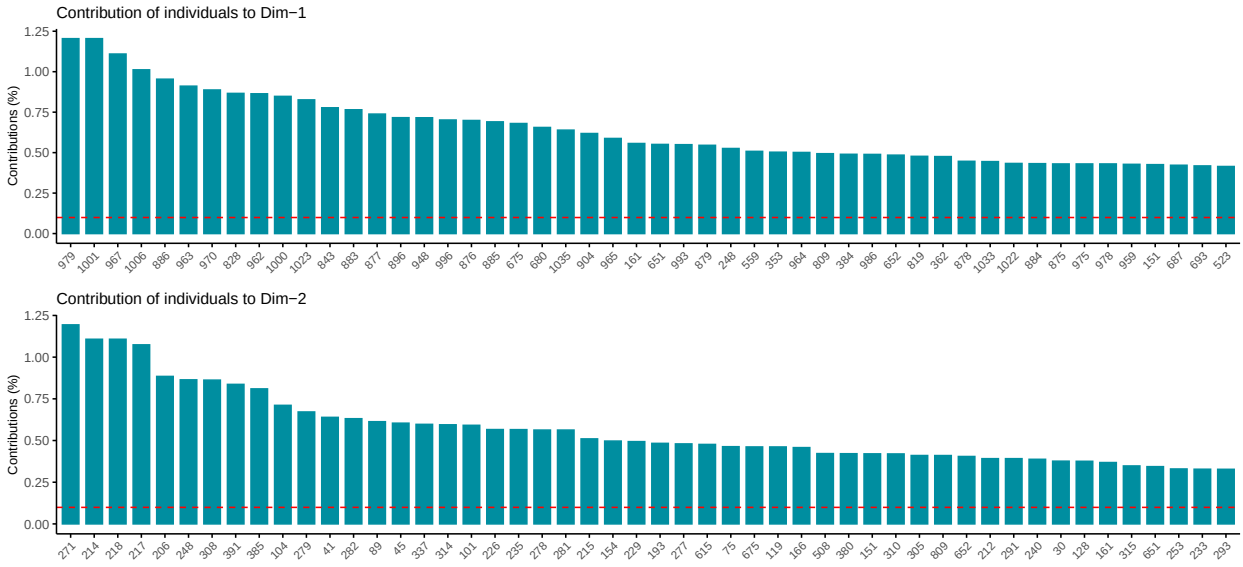


On remarque que les modalités qui contribuent le plus à l'axe 1 sont le fait de ne pas vouloir faire d'études supérieures (higher_no) et le fait d'être dans l'école Mousinho da Silveira (MS). Les deux autres modalités qui contribuent de façon assez conséquente sont le fait d'avoir redoublé 1 fois (failures_1) et le fait d'étudier moins de 2 heures par semaine (stud < 2). Ces modalités contribuent à presque 39% de l'axe 1.
Les modalités qui contribuent le plus à l'axe 2 sont les mathématiques (et le portugais mais moins car les

notes sont meilleures), le fait d'être absent plus de 10 fois (abs]10,75]), le fait de prendre des cours particuliers (paid_yes), le fait de redoubler 1 fois (failures_1) et le fait d'avoir une moyenne entre 5 et 10 (moy]5,10]). Ces modalités contribuent à 60% de l'axe 2.

4.1.2 Individus

On regarde quels individus contribuent le plus :



On regarde maintenant les individus qui contribuent le plus à chaque axe :

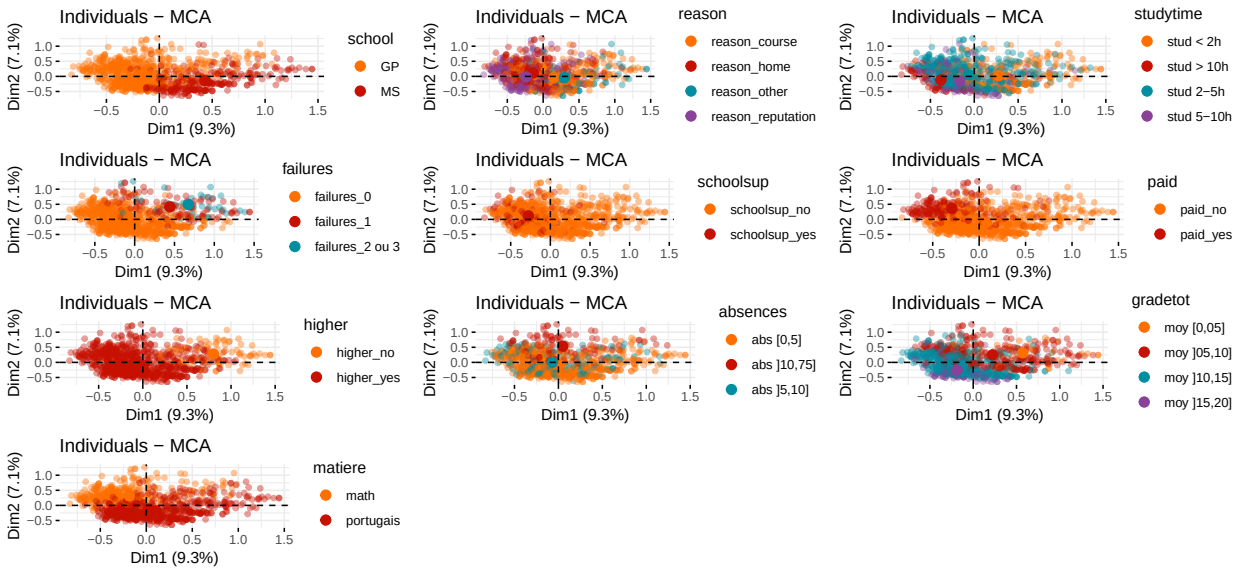
Table 14 – Individus qui contribuent le plus à l axe 1

	school	reason	traveltime	studytime	failures	schoolsup	paid	nursery	higher	absences	gradetot	matiere	classe
979	MS	other	trav 15-30	stud < 2h	1	no	no	yes	no	abs [0,5]	moy [0,05]	portugais	12eme
1001	MS	other	trav 15-30	stud < 2h	1	no	no	yes	no	abs [0,5]	moy [0,05]	portugais	12eme
967	MS	course	trav 15-30	stud < 2h	2 ou 3	no	no	no	no	abs [0,5]	moy]05,10]	portugais	11eme
1006	MS	course	trav 15-30	stud 2-5h	2 ou 3	no	no	yes	no	abs [0,5]	moy [0,05]	portugais	11eme

Table 15 – Individus qui contribuent le plus à l axe 2

	school	reason	traveltime	studytime	failures	schoolsup	paid	nursery	higher	absences	gradetot	matiere	classe
271	GP	home	trav < 15	stud 2-5h	2 ou 3	no	yes	yes	yes	abs]10,75]	moy]05,10]	math	12eme
214	GP	home	trav < 15	stud 2-5h	1	no	yes	yes	yes	abs]10,75]	moy]05,10]	math	12eme
217	GP	reputation	trav < 15	stud 2-5h	2 ou 3	no	yes	yes	yes	abs]10,75]	moy]05,10]	math	10eme
218	GP	home	trav < 15	stud 2-5h	1	no	yes	yes	yes	abs]10,75]	moy]05,10]	math	12eme

Les étudiants 979, 1001 et 967 contribuent le plus à l'axe 1 car ils étudient moins de 2 heures par semaine et ont une moyenne entre 0 et 5 tout en ayant redoublé au moins une fois, en étant dans l'école Mousinho da Silveira et ne voulant pas faire d'études supérieures (ils n'aiment très certainement pas l'école). L'étudiante 1006 comporte les mêmes caractéristiques, sauf qu'elle travaille de 2 à 5 heures par semaine. Étant donné que l'on est sur la matière portugais, on peut penser que soit cette étudiante a menti sur le temps de travail par semaine qu'elle fait, soit qu'elle est dyslexique car son niveau de santé est très bas, ce qui peut l'handicaper dans ce cours.
Du côté de l'axe 2, les individus 271, 214, 217 et 218 sont surtout ceux qui sont souvent absents (plus de 10 fois), dans la matière des mathématiques tout en ayant redoublé au moins une fois en prenant des cours particuliers et en travaillant de 2 à 5 heures par semaine. Il s'agit très certainement d'étudiants en difficultés ou alors qui n'aiment pas les mathématiques mais qui sont forcés par leurs parents à en faire.



— Variable school : l'école Mousinho da Silveira représente toute la partie en bas à droite au niveau de l'axe 1.
— Variable reason : les raisons liées au domicile proche de l'école et la réputation sont plutôt au niveau des bonnes notes.
— Variable studytime : les étudiants qui travaillent plus de 5 heures par semaine se situent au niveau des bonnes notes tandis que les étudiants qui travaillent moins de 2 heures par semaine se situent au niveau des mauvaises notes.
— Variable failures : les étudiants ayant redoublé au moins une fois se situent au niveau des mauvaises notes en haut à droite.
— Variable schoolsup : l'aide vient surtout au niveau de la matière mathématiques en haut à gauche.
— Variable paid : les étudiants ayant payé des cours personnels se retrouvent en haut à gauche au niveau de l'axe 2.
— Variable higher : les étudiants qui ne veulent pas faire d'études supérieures sont représentés sur l'axe 1 (au niveau des redoublements et des mauvaises notes).
— Variable absences : comme vu précédemment, on retrouve les individus les plus absents sur l'axe 2. Nous remarquons également que les étudiants ayant redoublé au moins une fois ne sont pas forcément beaucoup absents.
— Variable gradetot : on remarque que la moyenne semble avoir une relation croissante entre les axes 1 et 2, c'est-à-dire qu'à mesure que l'on augmente la valeur dans l'axe 1 et 2, les notes descendent.
— Variable matiere : il semblerait que l'axe des abscisses les coupe en deux avec le portugais en bas et les mathématiques en haut. Cependant, on remarque que les mauvaises notes se situent beaucoup en portugais (ce qui peut être lié aux différences de taille d'échantillon entre les deux matières), tout comme les très bonnes.
— Pour terminer, les variables traveltime, nursery et classe n'apportaient rien donc nous ne les avons pas mises.

4.1.3 Récapitulatif

Les axes 1 et 2 représentent les variables qui font que les notes sont mauvaises, les facteurs de non réussite à l'école (redoublements, moyenne en dessous de la moyenne, veut pas faire d'études supérieures, étudiants absents, mathématiques qui sont moins réussis que le portugais).

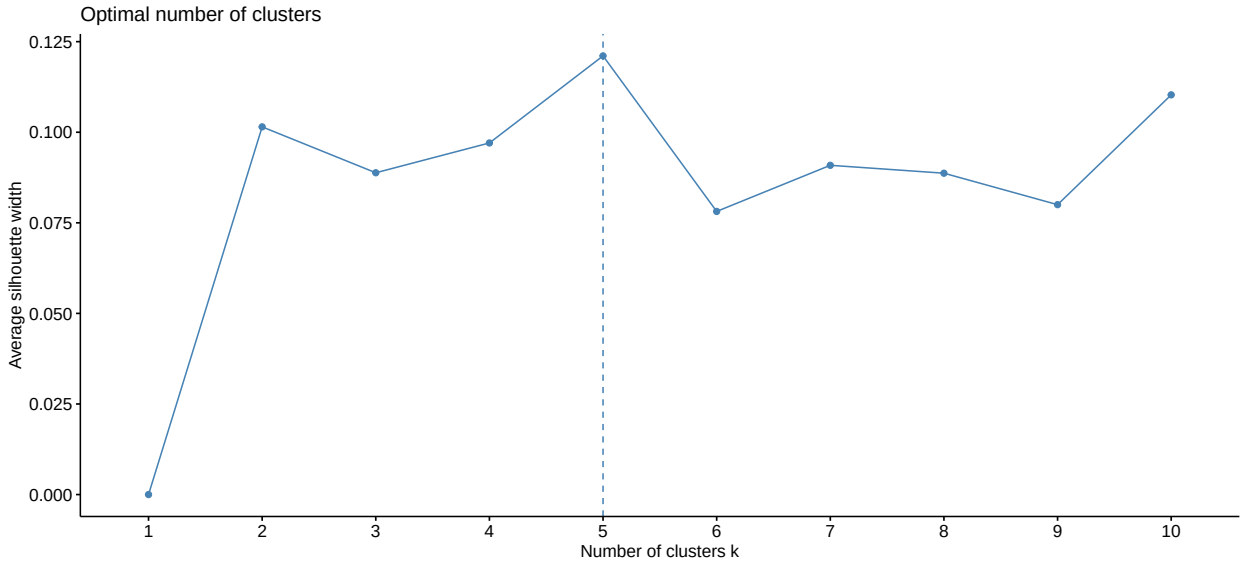
5 Classification

5.1 Environnement familial

5.1.1 K-means

Nous utilisons tout d'abord la méthode des k-means qui est une technique de la méthode de partitionnement.

Nous cherchons le nombre optimal de cluster proposé par cette méthode :



D'après la méthode silhouette dans la méthode de partitionnement des k-means, le nombre de clusters optimal serait de 5.

5.1.2 CAH

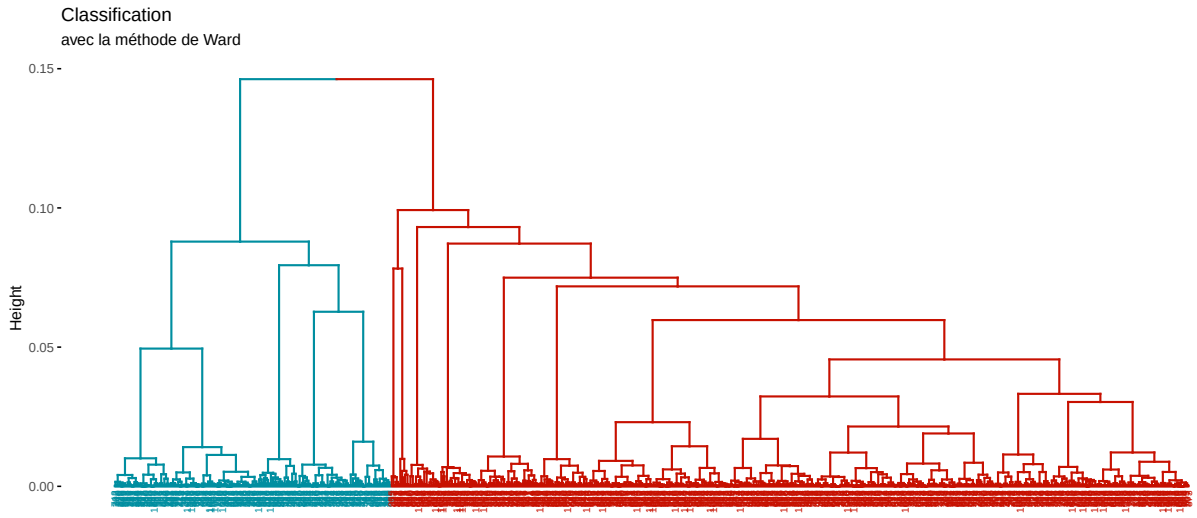
À la suite de la méthode par partitionnement, nous faisons un classification hiérarchique pour comparer les deux méthodes :

Tout d'abord, voici un tableau qui présente le gain d'inertie entre la CAH avec et sans consolidation.

TABLE 16 – Gain d'inertie après consolidation

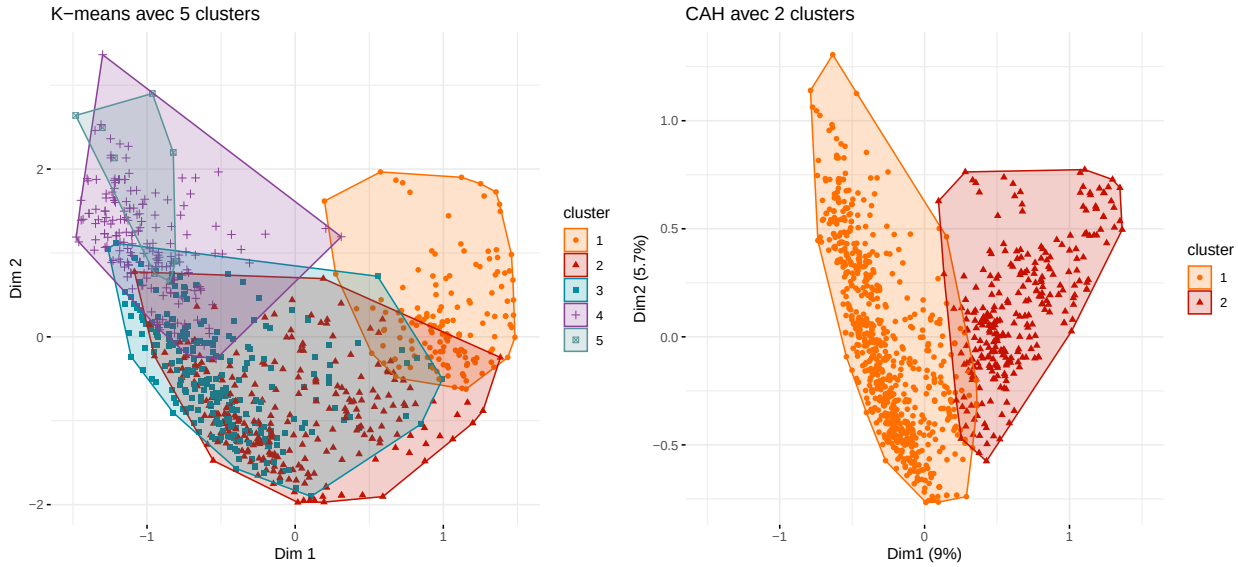
	before	after
variance	0.146	0.185

Nous pouvons constater une augmentation de l'inertie de 0.04 lorsque nous passons de sans à avec consolidation.



En observant le dendogramme, nous pouvons dire que les 5 clusters proposés par la méthode de partitionnement ne semblent pas vraiment être en adéquation avec le dendogramme présent ci-dessus où l'on peut estimer que 2 clusters serait le meilleur choix.

Pour appuyer nos propos, comparons les représentations des clusters pour la méthode des k-means et la méthode CAH.



Il semble clairement que la méthode des k-means n'est pas optimale, les clusters se superposent et donc le graphique n'est pas lisible, on n'arrive pas à degager 2 clusters distinctement. Si on utilise la CAH, on constate bien 2 clusters bien différents qui semblent être coupés par un axe vertical.

5.1.3 Modalités des variables pseudo-test de proportions

Voici le tableau représentant les modalités des variables les plus présentes dans le premier cluster.

Table 17 – Cluster 1

	Cla/Mod	Mod/Cla	Global	v.test	p.value
Medu=Medu_niveau 2	96.47	38.13	28.05	12.52	6.153e-36
Fedu=Fedu_niveau 1	98.34	33.10	23.89	12.47	1.130e-35
Mjob=Mjob_other	91.73	49.58	38.35	12.16	5.112e-34
Medu=Medu_niveau 1	99.47	26.40	18.83	11.53	9.686e-31
Medu=Medu_niveau 3	94.67	29.75	22.30	9.88	4.851e-23
Mjob=Mjob_at_home	94.18	24.86	18.73	8.67	4.141e-18
Fedu=Fedu_niveau 2	88.46	38.55	30.92	8.63	6.300e-18

Interprétation pour le cluster 1 :

— 96% de la modalité Medu_niveau2 est présente dans le cluster et représente 38% des individus dans celui-ci.
— Si on regarde Medu_niveau1, Medu_niveau2 et Medu_niveau3, les trois représentent 93% des individus de ce cluster : on peut dire que dans ce cluster, les étudiants dont les mères ont fait très peu d'études sont majoritaires.
— Pour le niveau d'étude du père, la part d'étudiants dont le père a fait très peu d'études dans ce cluster monte jusqu'à 71%.
— La SCP de la mère joue égalament un rôle dans la création du cluster car si on additionne Mjob_other et Mjob_at_home, 75% des étudiants de ce cluster prennent ces 2 modalités.

En conclusion du cluster 1, il représente les étudiants dont les parents ont un faible niveau d'études et dont la mère travaille à la maison ou prennant la modalité others.

Table 18 – Cluster 2

	Cla/Mod	Mod/Cla	Global	v.test	p.value
Medu=Medu_niveau 4	89.40	92.15	29.93	28.37	5.164e-177
Fedu=Fedu_niveau 4	81.28	60.75	21.70	18.63	1.747e-77
Mjob=Mjob_teacher	96.15	42.66	12.88	17.68	5.968e-70
Fjob=Fjob_teacher	96.83	20.82	6.24	11.88	1.514e-32
Mjob=Mjob_health	83.12	21.84	7.63	10.26	1.067e-24
failures=failures_0	32.87	96.59	85.33	7.15	8.410e-13
Fjob=Fjob_health	80.00	10.92	3.96	6.76	1.400e-11

Interprétation cluster 2:

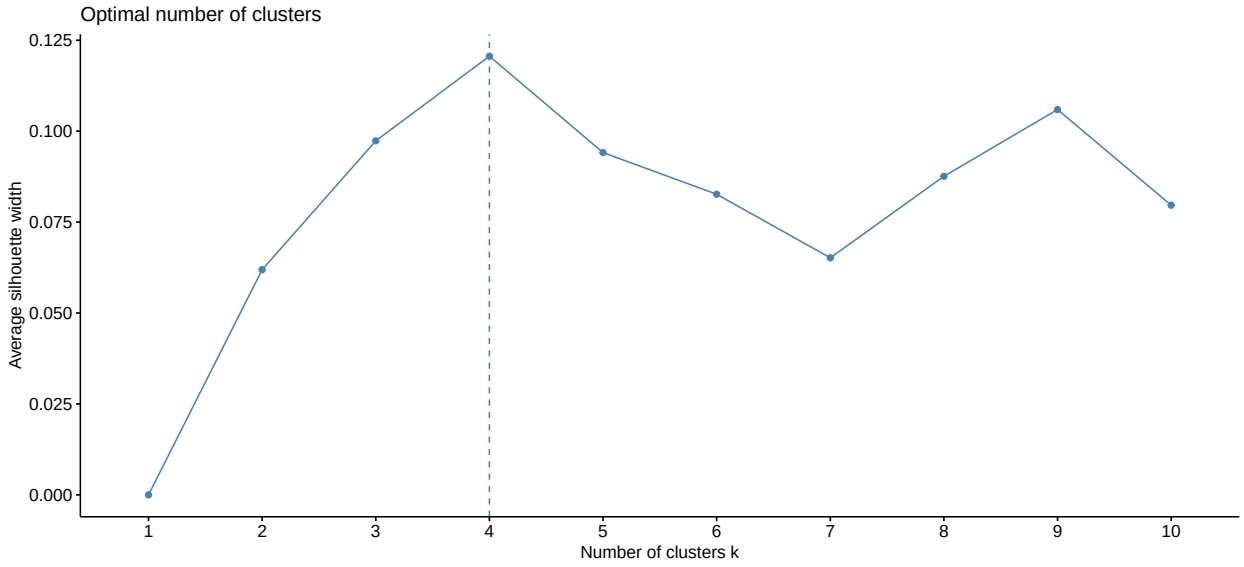
— La modalité Medu_niveau4 est présente pour près de 92% des individus de ce cluster. Les 4 autres modalités de cette variable représentent donc que 9%.
— Pour la modalité Fedu_niveau4, 81% des individus de cette modalité sont dans ce cluster et 60% des individus de ce cluster ont un père qui a un niveau d'éducation très élevé.
— Pour les CSP, 96% des mères et pères professeurs sont situés dans ce second cluster.
— Également concernant les SCP, entre 80 et 83% des mères et pères dans le système de santé font partie de ce cluster.

En conclusion, on voit clairement que l'étudiant type de ce cluster correspond à l'étudiant dont les parents ont un haut niveau d'études. Ce qui fait que les étudiants dont les parents travaillent dans les secteurs de l'éducation et de la santé sont très présents.

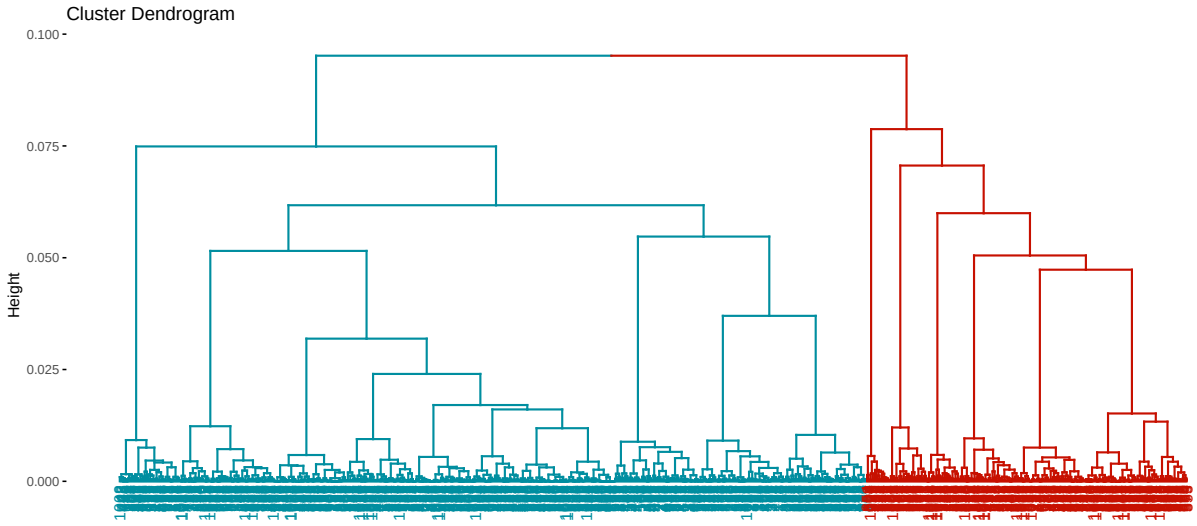
Finalement, grâce à ces deux tableaux, on retrouve bien les résultats obtenus lors de l'ACM où le premier axe séparait le niveau d'éducation mais également les CSP des parents des étudiants. De plus, grâce à l'AFC réalisée sur les CSP on peut en conclure que les deux clusters séparent les étudiants ayant en général une bonne note et étant fils de professeurs ou de médecins et les étudiants ayant une moins bonne note et étant fils de parents qui travaillent à la maison ou dans la catégorie "other".

5.2 Environnement scolaire

Notre objectif est de faire une classification non supervisée sur notre ACM sur les variables du milieu scolaire. On décide de garder les 20 premiers axes car ils comprennent 91% de l'information totale. Par la méthode des k-means on détermine une première découpe :



Le nombre de cluster optimal serait donc de 4. On regarde notre dendrogramme pour voir si cela pourrait être cohérent :



Ici, il semblerait plutôt que l'on ai 2 clusters bien distinct. Nous allons donc faire notre classification ascendante hiérarchique avec 2 clusters.

Tout d'abord, voici un tableau qui présente le gain d'inertie entre la CAH avec et sans consolidation.

Table 19 – Gain d'inertie après consolidation

	before	after
variance	0.095	0.121

Nous pouvons constater une augmentation de l'inertie de 0.03 lorsque nous passons de sans à avec consolidation.

On va regarder quels sont les axes qui représentent le mieux nos clusters afin de faire une représentation graphique :

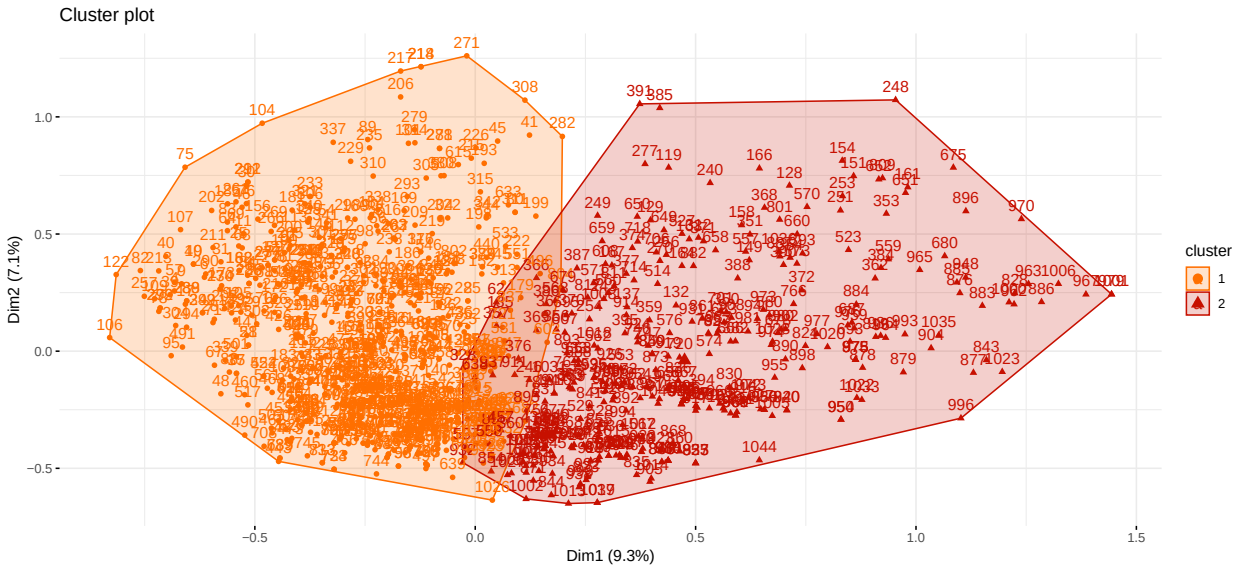
Table 20 – Les axes qui représentent le mieux le cluster 1

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.2	2.424	0.020	0	0.362	0.363	0.015
Dim.6	2.373	0.016	0	0.255	0.299	0.018
Dim.12	-2.214	-0.014	0	0.248	0.275	0.027
Dim.4	-3.362	-0.023	0	0.344	0.309	0.001
Dim.7	-3.465	-0.023	0	0.252	0.296	0.001
Dim.5	-3.920	-0.027	0	0.286	0.306	0.000
Dim.1	-26.010	-0.242	0	0.194	0.414	0.000

Table 21 – Les axes qui représentent le mieux le cluster 2

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.1	26.010	0.475	0	0.306	0.414	0.000
Dim.5	3.920	0.053	0	0.336	0.306	0.000
Dim.7	3.465	0.045	0	0.363	0.296	0.001
Dim.4	3.362	0.046	0	0.217	0.309	0.001
Dim.12	2.214	0.027	0	0.319	0.275	0.027
Dim.6	-2.373	-0.031	0	0.368	0.299	0.018
Dim.2	-2.424	-0.039	0	0.363	0.363	0.015

L'amas 1 est le mieux représenté par la dimension 2 et l'amas 2 est le mieux représenté par la dimension 1. Nous allons donc faire un graphique pour représenter les clusters sur les deux premiers axes :



Nous allons maintenant regarder les modalités pour chaque amas afin de comprendre ce que représente nos deux clusters. Voici un premier tableau sur les modalités du cluster 1 :

Table 22 – Cluster 1

	Cla/Mod	Mod/Cla	Global	v.test	p.value
school=GP	85.60	95.21	73.64	21.92	1.751e-106
higher=higher_yes	72.19	99.85	91.58	13.55	8.050e-42
traveltime=trav < 15	82.43	75.15	60.36	13.45	2.882e-41
failures=failures_0	72.82	93.86	85.33	10.39	2.671e-25
address=U	75.44	83.23	73.04	10.01	1.399e-23
matiere=math	82.89	46.41	37.07	8.84	9.682e-19
Medu=Medu_niveau 4	84.44	38.17	29.93	8.32	8.782e-17

La modalité la mieux représentée correspond à l'école Gabriel Pereira. On remarque que 95% des étudiants du cluster sont de l'école Gabriel Pereira. De plus, les modalités higher_yes, trav < 15, failures_0 et U sont surreprésentés. Ainsi, 99.8% des étudiants du cluster veulent faire des études supérieures, 75% habitent à moins de 15 minutes, 94% n'ont jamais redoublé et 83.3% vivent en ville. Nous remarquons donc que les étudiants de ce groupe sont plutôt de bons étudiants (ne redoublent pas, veulent faire des études supérieures), qui habitent en ville (donc un temps de trajet plus court). Nous allons maintenant regarder les modalités du deuxième cluster :

Table 23 – Cluster 2

	Cla/Mod	Mod/Cla	Global	v.test	p.value
school=MS	87.97	68.62	26.36	21.92	1.751e-106
higher=higher_no	98.82	24.63	8.42	13.55	8.050e-42
studytime=stud < 2h	58.86	51.61	29.63	10.74	6.361e-27
address=R	58.82	46.92	26.96	10.01	1.399e-23
traveltime=trav 30-1h	86.49	18.77	7.33	9.72	2.438e-22
Medu=Medu_niveau 1	63.16	35.19	18.83	9.23	2.735e-20
gradetot=moy]05,10]	54.22	48.97	30.53	8.95	3.584e-19

La modalité la mieux représentée est l'école Mousinho da Silveira. On remarque que 68.6% des étudiants de ce cluster sont dans l'école alors que cette modalité ne représente que 26% de la variable. De plus, les autres modalités qui représentent le mieux ce cluster sont higher_no, stud < 2h, R, trav 30-1h. Cependant, toutes ces modalités représentent une faible proportion de la variable (respectivement : 8%, 30%, 27% et 7%) mais on remarque que dans ce cluster, on retrouve 99% des étudiants ne voulant pas faire d'études supérieures, 59% des étudiants travaillant moins de 2 heures, 59% des étudiants vivant à la campagne et 86.5% des étudiants qui ont un trajet qui dure entre 30 minutes et 1 heure. Cet amas représente donc plutôt les mauvais étudiants (ceux qui ne veulent pas faire d'études supérieures, qui travaillent moins de 2 heures par semaine) et les étudiants qui habitent à la campagne (et par conséquent, qui ont des trajets plus longs).

En conclusion, notre classification ascendante hiérarchique a construit 2 clusters, l'un correspondant à l'école Mousinho da Silveira et l'autre correspondant à l'école Gabriel Pereira, ce qui nous montre les différences entre les 2 écoles. Il nous semblerait donc que les étudiants de l'école Mousinho da Silveira sont plutôt de mauvais étudiants qui vivent en campagne et les étudiants de l'école Gabriel Pereira sont plutôt de bons étudiants vivant en ville. Nous sommes donc peut-être en train d'observer les inégalités d'éducation au Portugal entre les étudiants habitant à la ville et ceux habitant à la campagne.

6 Conclusion

Après avoir décidé de séparer notre base de données en trois axes de travail, nous avons pu constater l'importance de nombreuses variables pour expliquer la performance des étudiants portugais.

1. Pour l'environnement familial, le niveau d'éducation et les catégories socioprofessionnelles des parents ont un effet sur la performance de leur enfant. Plus le niveau d'éducation des parents est élevé ou plus ils ont une CSP élevée, plus la moyenne de l'élève semble s'améliorer.

2. Pour l'environnement extra-scolaire, nous avons pu observer que le niveau de temps libre, de sortie avec des amis et la consommation d'alcool avait un effet négatif sur la performance des étudiants si le niveau de ces trois variables était élevé.

3. Pour l'environnement scolaire, le nombre de redoublement, le temps passé à étudier et la matière sont des facteurs importants dans les performances des étudiants portugais. On constate également que l'école et donc le lieu d'habitation sont très important dans la réussite des étudiants.